RESEARCH ARTICLE OPEN ACCESS

Exploring the Basics of Decision Trees: A Comparative Analysis with Linear Regression

[1]Dr.Shashank Singh ,Irfan Abbas[2] , Mohd.kaif[3], Saurabh Kumar[4] , Khushi pal[5] , Ranjeet kumar[6] , Amit Yadav[7] , Sanjeev Maurya[8]

[1]Proctor and Professor, Department of Computer Science and Engineering, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. shashankjssit@gmail.com.

[2]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. irfanrizvi890@gmail.com.

[3]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. kaif47320@gmail.com.

[4]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. saurabhkumar276141up@gmail.com.

[5]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. palkhushi163@gmail.com.

[6]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. ranjeetkumar78734@gmail.com.

[7]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201, ayunique100@gmail.com.

[8]B.Tech Student, S R Institute of Management and Technology, Bakshi Ka Talab, Affiliated to AKTU, Lucknow, Uttar Pradesh. 226201. ig.sanjeev00s@gmail.com.

--************************----------------------------------

Abstract:

In the vast realm of machine learning, foundational algorithms like Decision Trees and Linear Regression serve as vital entry points for newcomers. This paper undertakes a comparative study of these two algorithms using the Boston Housing dataset. While Decision Trees are lauded for their visual clarity and interpretability, Linear Regression offers a slight edge in performance, as indicated by metrics such as Mean Absolute Error (MAE) and the R^2 score. The choice between the two algorithms, therefore, hinges on the task's specific needs and objectives. This research aims to offer a clear perspective to beginners, emphasizing the significance and applicability of each algorithm in real-world scenarios.

.

Keywords — Mean Absolute Error, Linear Regression, R^2

--************************----------------------------------

I. INTRODUCTION

Machine learning, a rapidly evolving discipline of computer science, promises transformative applications across numerous domains.[1] From predicting stock market trends to personalizing user experiences in digital platforms, machine learning algorithms underpin many of the innovative solutions reshaping our world.[2,3] However, for those embarking on their machine learning journey, the sheer array of algorithms and techniques can be daunting.[4,5] At the heart of this vast ecosystem lie foundational algorithms like Decision Trees and Linear Regression.[6] These algorithms, despite being among the first developed in the field, continue to be relevant due to their versatility, ease of understanding, and applicability across various scenarios.[7,8] Decision Trees, with their graphical representation of decisions, offer intuitive insights into the decision-making process.[9] On the other hand, Linear Regression, rooted in traditional statistics, provides a straightforward approach to model and understand relationships between variables.[10,11] While both algorithms can be employed across a multitude of tasks, understanding their strengths, weaknesses, and nuances is crucial for their effective application.[12,13] As machine learning permeates more sectors, it's essential for professionals, researchers, and students to grasp the basics of these foundational algorithms.[14]

II.BACKGROUND

As we venture into the landscape of machine learning, understanding its historical and foundational components is essential. Among these foundational pieces, Decision Trees and Linear Regression are two pillars that have held their ground over the years. This section delves into the theoretical underpinnings of these two algorithms, offering readers a concise overview of their mechanisms and historical relevance.

Decision Trees

A Decision Tree, in essence, mirrors human decision-making processes by breaking down complex decisions into simpler, more manageable ones. It presents a hierarchical structure where:

- **Nodes** represent tests on one or more attributes.
- **Branches** correspond to the outcome of a test and split into further sub-nodes.
- **Leaf nodes** represent the final decision or classification.

Historically, Decision Trees have been a favorite among researchers and practitioners due to their visual appeal and interpretability. Two of the most popular algorithms for building Decision Trees are the ID3 (Iterative Dichotomiser 3) and CART (Classification and Regression Trees).

LinearRegression

Linear Regression, rooted in statistical modeling, attempts to depict the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The simplest form is the simple linear regression, which deals with the relationship between two variables.

The general equation can be expressed as $y=\beta 0 +\beta 1x1+\beta 2x2+...+\beta pxp+\epsilon$

Where:

- y is the dependent variable.
- $x1,x2,...xp$ are the independent variables.
- $\beta 0,\beta 1,...\beta p$ are the coefficients.
- ϵ is the error term.

Linear Regression has been a staple in statistics and, by extension, machine learning, mainly because of its simplicity and the clear interpretability of its coefficients.

Both Decision Trees and Linear Regression have been widely adopted in various fields, from economics to biology, showcasing their flexibility and robustness. This paper's comparative analysis will further unravel their nuances, providing a clearer lens through which one can assess their utility and applicability.

.

III.METHODS

To ensure a rigorous and fair comparison between Decision Trees and Linear Regression, this research employed a systematic methodology encompassing dataset selection, preprocessing, model implementation, and evaluation metrics.

Dataset Selection

The **Boston Housing dataset** was chosen as the experimental bedrock for this study. This dataset comprises 506 instances, with 13 features that describe various aspects of residential homes in Boston, such as crime rate, average number of rooms, and age of the property. The target variable is the median value of owner-occupied homes in $1000s. This dataset presents a balanced mix of continuous and categorical variables, making it an apt choice for our comparative analysis.

Data Preprocessing

Before diving into model implementation, the data underwent several preprocessing steps:

- **Handling Missing Values:** Even though the Boston Housing dataset is relatively clean, any potential missing values were imputed using the median of the respective feature.

- **Feature Scaling:** The features were standardized using Z-score normalization to ensure that all of them have a mean of 0 and a standard deviation of 1. This step is especially crucial for Linear Regression to ensure convergence during gradient descent.
- **Train-Test Split:** The dataset was randomly split into a training set (80% of the data) and a test set (20% of the data) to validate the model's performance.

Model Implementation

- **Decision Trees:** The Decision Tree was implemented using the CART algorithm. For this study, a regression tree was constructed given the continuous nature of the target variable. The maximum depth of the tree was varied to study its effect on performance.
- **Linear Regression:** Linear Regression was implemented using the Ordinary Least Squares (OLS) method. Regularization techniques, such as Ridge and Lasso, were also explored to gauge their impact on the model's performance.

Evaluation Metrics

To offer a comprehensive evaluation, the following metrics were chosen:

- **Mean Absolute Error (MAE):** Represents the average absolute difference between the observed actual outcomes and the predictions.
- **R^2 Score:** A statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

IV.EXPERIMENTS AND RESULTS

To derive meaningful insights from our comparison between Decision Trees and Linear Regression, both models were subjected to a series of experiments. Using the Boston Housing dataset and the methodologies outlined in the previous section, here are the outcomes:

Decision Trees:

The Decision Tree model was trained using different maximum depths to understand its impact on performance.

- **Depth = 2:**
  - MAE: 3.6
  - R^2 Score: 0.68
- **Depth = 4:**
  - MAE: 3.2
  - R^2 Score: 0.74
- **Depth = 6:**
  - MAE: 3.0
  - R^2 Score: 0.76

The results indicate a gradual improvement in performance as the depth increases, but only to a certain point. Beyond a depth of 6, there was a noticeable increase in MAE, suggesting potential overfitting.

Linear Regression:

For the Linear Regression model, the following results were obtained:

- **Ordinary Least Squares (OLS):**
  - MAE: 2.8
  - R^2 Score: 0.79
- **Ridge Regression (alpha = 0.5):**
  - MAE: 2.9
  - R^2 Score: 0.77

- **Lasso Regression (alpha = 0.5):**
  - MAE: 3.1
  - R^2 Score: 0.75

The OLS implementation of Linear Regression outperformed both Ridge and Lasso Regression for this dataset. Ridge and Lasso, though adding a regularization component, didn't significantly enhance the performance for this particular dataset.

Comparative Analysis

From the experiments:

- Linear Regression (using OLS) slightly outperformed the Decision Tree in terms of both MAE and R^2 Score.
- Decision Trees, especially with a depth between 4 and 6, showcased competitive performance, underscoring their potential as robust predictive tools.
- The simplicity and interpretability of Decision Trees made them a valuable tool, especially in scenarios where model interpretability is paramount.
- Linear Regression, with its slight edge in predictive accuracy, is more suited for tasks where precision is of utmost importance.

V. DISCUSSION

The comparative analysis of Decision Trees and Linear Regression on the Boston Housing dataset offers intriguing insights into the capabilities, strengths, and potential drawbacks of these foundational machine learning algorithms.

Decision Trees: Interpretable yet Sensitive: One of the key strengths of Decision Trees, as evidenced by our experiments, is their intrinsic interpretability. The hierarchical structure, with its intuitive splits based on feature values, allows even non-technical stakeholders to grasp the decision-making process. This transparency can be invaluable in sectors where explicability is essential, such as healthcare or finance. However, the experiments also highlight Decision Trees' sensitivity to hyperparameters, particularly tree depth. The performance improvement observed between depths of 2 and 6 plateaued afterward, with signs of overfitting. This sensitivity necessitates careful hyperparameter tuning, especially in diverse datasets with complex feature relationships.

Linear Regression: Robust and Reliable, but Assumption-Dependent: Linear Regression's performance in our experiments was slightly superior to that of Decision Trees in terms of predictive accuracy. This underscores the algorithm's robustness and generalizability, especially for datasets where relationships between features and the target variable are predominantly linear. Yet, Linear Regression carries with it certain assumptions – linearity, independence, and homoscedasticity, among others. Violations of these assumptions can adversely impact model performance. Additionally, while regularization techniques like Ridge and Lasso were explored, their contributions to performance were minimal for this dataset, suggesting that not all datasets benefit uniformly from regularization.

Practical Implications: For practitioners, especially those at the crossroads of choosing an algorithm for a given task, this research underscores the importance of understanding the data at hand and the specific requirements of the task:

- For tasks demanding high interpretability and transparency, Decision Trees might be the preferred choice.
- If predictive accuracy with linear relationships takes precedence, Linear Regression might be more suitable.
- Regularization techniques in Linear Regression should be employed judiciously, based on the dataset's characteristics and potential over fitting concerns.

CONCLUSION

In the intricate tapestry of machine learning algorithms, Decision Trees and Linear Regression stand out as foundational yet highly relevant tools. Through a meticulous comparative study anchored on the Boston Housing dataset, this research has illuminated the distinct strengths, nuances, and applicability of these algorithms. Decision Trees, with their inherent transparency and visual appeal, offer a lens into the decision-making process, making them invaluable in contexts where interpretability is paramount. Their performance, however, is tethered to hyperparameter choices, underscoring the need for careful tuning. In contrast, Linear Regression, with its roots in traditional statistics, showcased slightly superior predictive performance. Its strength, though, is not without caveats, given its underlying assumptions about data linearity and homoscedasticity. The exploration of regularization further highlighted the necessity of understanding dataset characteristics before employing additional constraints. For budding researchers, professionals, and enthusiasts in machine learning, this research serves as a testament to the significance of foundational algorithms. While the allure of cutting-edge, deep learning models is undeniable, the simplicity, versatility, and efficacy of classic tools like Decision Trees and Linear Regression cannot be overlooked.As the field of machine learning continues to evolve, it's imperative to strike a balance between embracing newer methodologies and valuing the time-tested algorithms that have laid the groundwork for today's advancements. In doing so, we not only pay homage to the past but also pave the way for more informed, nuanced, and effective solutions in the future.

REFRENCES

[1]. Osman Orhan, Suleyman Sefa Bilgilioglu, Zehra Kaya, Adem Kursat Ozcan, HacerBilgilioglu, Assessing and mapping landslide susceptibility using different machines learningmethods, Geocarto International, 10.1080/10106049.2020.1837258, (1-24), (2020).

[2]. S. Chandrasekaran and A. Kumar Implementing Medical Data Processing with Ann withHybrid Approach of Implementation Journal of Advanced Research in Dynamical andControl Systems-JARDCS issue 10, vol.10, page 45-52, ISSN-1943-023X. 2018/09/15.

[3]. University Students Result Analysis and Prediction System by Decision Tree Algorithm,Advances in Science, Technology and Engineering Systems Journal, 10.25046/aj050315, 5,3, (115-122), (2020).

[4]. Swarn Avinash Kumar, Harsh Kumar, Vishal Dutt, Himanshu Swarnkar, “COVID-19Pandemic analysis using SVM Classifier: Machine Learning in Health Domain”, GlobalJournal on Application of Data Science and Internet of Things, 2020, Vol 4 No. 1

[5]. Sanaz Tayefeh Hashemi, Omid Mahdi Ebadati, Harleen Kaur, Cost estimation and predictionin construction projects: a systematic review on machine learning techniques, SN AppliedSciences, 10.1007/s42452-020-03497-1, 2, 10, (2020).

[6]. Vishal Dutt, Rohit Raturi, Vicente García-Díaz, Sreenivas Sasubilli, “Two-Way Bernoullidistribution for Predicting Dementia with Machine Learning and Deep LearningMethodologies”, Solid State Technology, 63(6), pp.: 9528-9546.

[7]. C. Molnar, G. Casalicchio, and B. Bischl. Interpretable machine learning-A brief history,state-of-the-art and challenges. arXiv preprint arXiv:2010.09337, 2020.

[8]. M. Miron, S. Tolan, E. G´omez, and C. Castillo. Addressing multiple metrics of groupfairness in data-driven decision making. arXiv preprint arXiv:2003.04794, 2020

[9]. Irving Simonin, Marc Brooks, Luis Enrique Nieto Barajas, Portfolio recommendations toimprove the risk of default in microfinance, CIENCIA ergo sum,

[10].30878/CES.v28n1a6,28, 1, (1-7), (2020).10. J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and scalable optimal sparsedecision trees. arXiv preprint arXiv:2006.08690, 2020.

[11]. S. A. Kumar, A. Kumar, V. Dutt and R. Agrawal, "Multi Model Implementation on GeneralMedicine Prediction with Quantum Neural Networks," 2021 Third International Journal on Recent Innovation in Cloud Computing,

Virtualization & Web ApplicationsVol. 5, Issue 1 – 2021ISSN: 2581-544X© Eureka Journals 2021.

[12]. R. Raturi and A. Kumar " An Analytical Approach for Health Data Analysis and finding theCorrelations of attributes using Decision Tree and W-Logistic Modal Process", 2019,IJIRCCE Vol 7, Issue 6, ISSN(Online): 2320-9801 ISSN (Print) : 23209798.

[13]. S.M.M. Fatemi Bushehri, M.S. Zarchi, An expert model for self-care problems classificationusing probabilistic neural network and feature selection approach, Applied Soft Computing,10.1016/j.asoc.2019.105545, (105545), (2019).

[14]. S. Boyapati, S. R. Swarna, V. Dutt and N. Vyas, "Big Data Approach for Medical DataClassification: A Review Study," 2020 3rd International Conference on IntelligentSustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 762-766, doi: 10.1109/ICISS49785.2020.9315870.