Statiscitcal Learning Methods Comparison

Jacob Thielemier

11 May 2024

Method Comparison

We are going to compare three statistical learning methods: Linear Regression, Bootstrap, and Extreme Gradient Boosting (XGBoost). Linear Regression, Bootstrap, and XGBoost are three different computational methods with unique use across the statistical and machine learning fields. Linear Regression is generally used for predictive modeling and understanding relationships between variables through the interpretation of a fitted linear model. Bootstrap focuses on statistical inference, providing estimates of parameters, their variances, and confidence intervals by resampling data with replacement. XGBoost, a post-2000 method, uses gradient boosting to build powerful predictive models and is known for its high performance and accuracy in classification and regression tasks. While Linear Regression and Bootstrap are used for fundamental statistical analysis and understanding variable relationships, XGBoost is used for creating efficient predictive models across many data types and sizes.

1 Linear Regression

Purpose: Linear regression is used primarily for predictive modeling and statistical analysis to understand the relationship between a dependent variable and one or more independent variables.

Mechanism: The method involves fitting a linear equation to observed data. The equation forms a line that best approximates the relationships between the dependent variable and each independent variable. This is usually achieved through the least squares method, which minimizes the sum of the squares of the residuals (the differences between observed values and the values predicted by the model).

Advantages: Simplicity and interpretability: Linear regression models are straightforward to understand and interpret, making them a good starting point for predictive analysis. Efficient and cost-effective: These models can be quickly and easily fit to data without needing intensive computational resources.

Disadvantages: Assumption-heavy: Linear regression assumes a linear relationship between the dependent and independent variables, normality in the errors, and homoscedasticity (constant variance of the errors). These assumptions don't always hold in real-world data, which can limit the model's applicability. Prone to outliers: Linear regression is sensitive to outlier values which can significantly affect the slope and intercept of the regression line.

Typical Use Cases: Linear regression is widely used in economics for demand/supply forecasting, in business for sales forecasting, in biology for dose-response modeling, and in many other fields wherever relationships between variables need to be elucidated and predictions made based on observed data.

2 Bootstrap

Purpose: Used to estimate the distribution of a statistic based on a sample data set, particularly effective in estimating mean, variance, bias, and confidence intervals.

Mechanism: Involves repeatedly sampling with replacement from the data set and calculating the statistic of interest for each sample to create an empirical distribution of the statistic.

Advantages: Does not require assumptions about the distribution of data (non-parametric). Useful for assessing the reliability of statistical estimates and for constructing hypothesis tests.

Disadvantages: Bootstrap methods can be computationally intensive as they require multiple resamplings to get accurate estimates. May not work well when the sample size is too small or the data are not representative of the population.

Typical Use Cases: Statistical inference, error estimation in complex estimators, model validation.

3 XGBoost

Purpose: A highly efficient and scalable implementation of gradient boosted trees designed for speed and performance.

Mechanism: Builds an ensemble of trees sequentially, with each new tree attempting to correct the errors made by the previous ones. Uses gradient descent to minimize a loss function when adding new models.

Advantages: Provides high predictive accuracy. Handles a variety of data types, quality, and distributions effectively. Includes built-in features for handling missing data and regularizing to prevent overfitting.

Disadvantages: More complex to understand and tune due to the numerous hyperparameters (e.g., number of trees, depth of trees, learning rate). Can overfit if not properly tuned or if data are too noisy.

Typical Use Cases: Classification and regression in areas ranging from finance to biology; competitive machine learning.

Dataset Description

Educational achievement is crucial for personal development and economic advancement. Recent research has highlighted various factors influencing student performance, including socio-economic background, study habits, and school environment. Students performance is very crucial in solving issues of the learning process and one of the important matters to measure learning outcomes. Alhazmi & Sheneamer (2023) The Student Performance dataset we selected is designed to examine the factors influencing academic student performance. The dataset consists of 10,000 student records with the outcome measure being a Performance Index rating of 0-100.

We will be using the following R packages from The Comprehensive R Archive Network (CRAN) to conduct our method comparison:

- knitr: A general-purpose package for dynamic report generation in R
- ggplot2: Create elegant data visualizations using the grammar of graphics
- dplyr: A grammar of data manipulation
- boot: Bootstrap functions (originally by Angelo Cantry for S)
- xgboost: Extreme gradient boosting
- readr: Read rectangular text data

Method Results

1 Linear Regression

We use our dataset to answer a specific question for statistical analysis. Our question is: How does the number of hours studied influence the students performance?

We are going to use simple linear regression to analyze how the covariate of interest, Hours.Studied, impacts the Performance.Index. We identify our confouders that are Sleep.Hours and Extracurricular.Activities. Our precision variables are Previous.Scores and Sample.Question.Papers.Practiced. We create a model to use to answer our question above.

- Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activities + Previous.Scores + Sample.Question.Papers.Practiced

Using R and our previously identified packages we can execute our model against the dataset to determine answer our question. We see from Table 1 below that each of our covariates are significant and have a P-value less than 0.01. This answers our question by defining that Hours.Studied is significant on the Performance.Index. We can see that for each hour studied the student performance increased by about 2.853 points.

Table 1: Linear Regression Summary

Term	Estimate	Std..Error	P.value	CI.2.5.	CI.97.5.
Intercept	-34.076	0.127	< 0.001	-34.325	-33.826
Hours Studied	2.853	0.008	< 0.001	2.838	2.868
Sleep Hours	0.481	0.012	< 0.001	0.457	0.504
Extracurricular Activities	0.613	0.041	< 0.001	0.533	0.693
Previous Scores	1.018	0.001	< 0.001	1.016	1.021
Sample Question Papers Practiced	0.194	0.007	< 0.001	0.180	0.208

2 Boostrap

We are going to use Bootstrap to estimate the accuracy of the linear regression model used above. We do that by interpreting the Bias which measures the average difference between the bootstrap estimates and the original estimate. A small bias suggests that your original estimate is close to the center of the bootstrap estimates, indicating stability and reliability in our estimations under resampling. We are going to conduct 1000 replications using the boot package in R. We can see from Table 2 below that the Bias for our Estimate is -0.00004 which is very small, suggesting that the bootstrap samples yield an intercept very close to the original estimate, indicating reliability. Our covariate of interest (Hours.Studied) is -0.00014 which is also very small showing the reliability of this predictor.

Table 2: Bootstrap Summary

Term	Original Est.	Bias	Standard Err.
Estimate	-34.07559	-0.00004	0.12574
Hours Studied	2.85298	-0.00014	0.00759
Sleep Hours	0.48056	-0.00027	0.01209
Extracurricular Activities	0.61290	-0.00200	0.04182
Previous Scores	1.01843	0.00008	0.00120
Sample Questions	0.19380	-0.00033	0.00692

3 XGBoost

As a new and efficient ensemble learning algorithm, XGBoost has been widely applied for its multitudinous advantages, but its classification effect in the case of data imbalance is often not ideal. Zhang et al. (2022) Looking at our code below we can see how we prepare our our data by selecting two matrices, one for

Hours.Studied and one for Performance.Index. Next we split our data using 80% for training and 20% for testing. Then we are used some general paramters to train the model.

```
dat1 <- read.csv("C:\\Users\\JThie\\OneDrive\\Desktop\\Spring 24\\Machine Learning\\Project\\Student_Pe

dat1$Extracurricular.Activities <- as.numeric(as.factor(dat1$Extracurricular.Activities))
data_matrix <- as.matrix(dat1 %>% select(Hours.Studied, Performance.Index))
labels <- dat1$Performance.Index

train_index <- sample(1:nrow(dat1), 0.8 * nrow(dat1))
train_data <- data_matrix[train_index,]
test_data <- data_matrix[-train_index,]
train_label <- labels[train_index]
test_label <- labels[-train_index]

params <- list(booster = "gbtree",
               objective = "reg:squarederror",
               eta = 0.1,
               max_depth = 6,
               min_child_weight = 1,
               subsample = 0.5,
               colsample_bytree = 0.5)
xgb_train <- xgb.DMatrix(data = train_data, label = train_label)
xgb_model <- xgb.train(params = params, data = xgb_train, nrounds = 100)

xgb_test <- xgb.DMatrix(data = test_data)
predictions <- predict(xgb_model, xgb_test)

rmse <- sqrt(mean((predictions - test_label)^2))
print(paste("RMSE:", rmse))
```

[1] "RMSE: 0.433011473631825"

Finally we make our predictions and evaluate the model we built. We can see above that our Root Mean Square Error (RMSE) is 0.721 which indicates that the average magnitude of the errors in our dataset is low. RMSE is a measure of the differences between values predicted by a model and the values actually observed from the environment that is being modeled. It is calculated by:

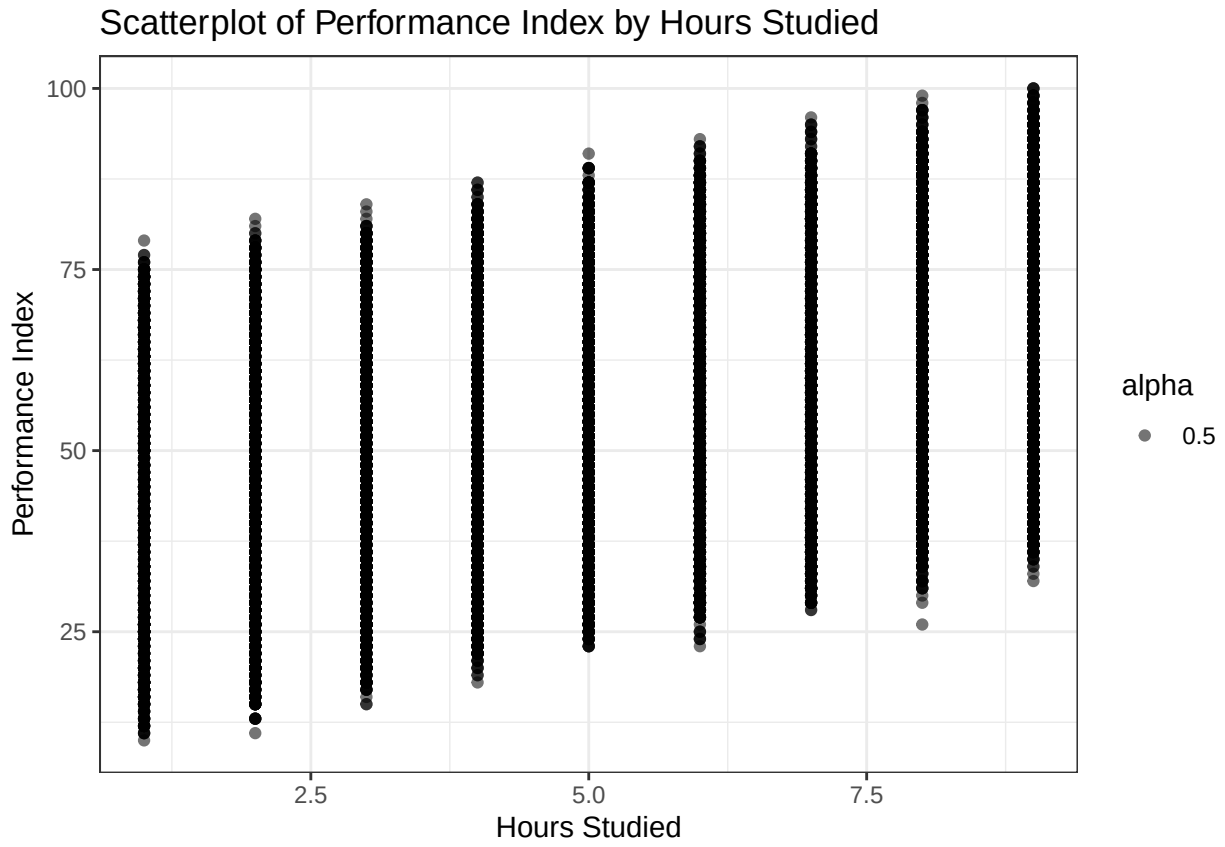
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$$

Our model is performing quite effective at predicting the Performance.Index based on the hours studied by the students.

Summary

We use the ggplot2 package to create a visual of our data. It can be easily interpreted from the plot below that student performance increases and hours studied also increases. Visualization of data can simplify Linear Regression analysis if the dataset meets the four linear regression assumptions of linearity, independence, normality, and equal variance. We then used Bootstrap to understand the accuracy of the linear regression model for each of the covariates. Lastly, we used XGBoost which gave us a clear understanding that our model performed well predicting the student performance using hours studied.

Overall, our three machine learning methods built upon one another to give us a holistic analysis of our Student Performance dataset. These models where not in direct competition which each other but served as different pieces of our interpretation and prediction of our data.



References

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3250702

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*. https://doi.org/10.1177/15501329221106935