Report on Development of Python software to address a supervised learning challenge

By: OJONUGWA WADA

INTRODUCTION

This report presents the development of Python software to address a supervised learning challenge. Data importing, visualization, dataset segmentation, model development, and regression model evaluation are among the primary activities. Predicting a continuous target variable ('y') from a collection of input characteristics (x1 to x13) is the main objective.

METHODOLOGY

The dataset was thoroughly analyzed to identify trends and connections. To observe how various traits relate to one another and change, visual aids were employed. The data was divided into two sections to create predictive models: 20% was used to assess the models' accuracy, while the remaining 80% was used for training.

To determine how effectively they predict results, two models Linear Regression and Random Forest Regressor were trained and evaluated. Their accuracy was compared using key performance indicators. To assess the models' dependability across several data samples, cross-validation was used. To determine the key elements affecting forecasts, a study was also carried out.

Interactive visuals were utilized to help explain the results. These made it easier to evaluate the two models' predictions and identify the most influential aspects.

Q&A Section

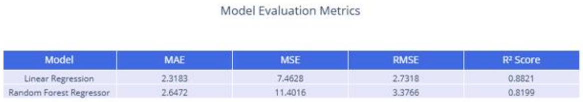
Question 1 (Python Code): What does the train_test_split function accomplish?

By separating the dataset into training and testing sections, train_test_split allows for objective model assessment through testing on unknown data.

While logistic regression uses a logistic function to describe the likelihood of categorical outcomes, making it perfect for classification tasks, linear regression assumes a linear connection between input characteristics and the target variable, making it appropriate for predicting continuous values.

RESULTS

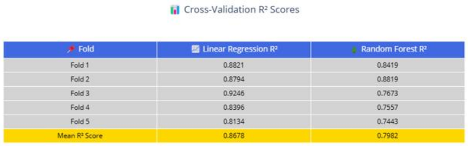
The evaluation's findings contrast the Random Forest Regressor and Linear Regression models. With lower MAE, MSE, and RMSE, Linear Regression performs better than Random Forest, suggesting higher prediction accuracy. The R^2 score provides additional evidence that the target variable's variation is better explained by linear regression.



Model Evaluation Metrics

Model	MAE	MSE	RMSE	R^2 Score
Linear Regression	2.3183	7.4628	2.7318	0.8821
Random Forest Regressor	2.6472	11.4016	3.3766	0.8199

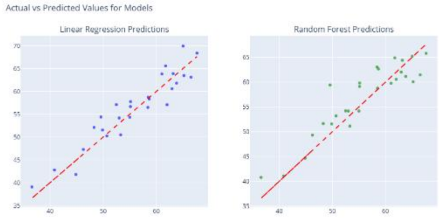
Cross-validation findings show that Linear Regression consistently produces higher mean R^2 scores across folds. The real vs. anticipated charts support this observation, with Linear Regression predictions closely following the ideal trend and Random Forest showing significant variation.



📊 Cross-Validation R^2 Scores

📌 Fold	📈 Linear Regression R^2	🌲 Random Forest R^2
Fold 1	0.8821	0.8419
Fold 2	0.8794	0.8819
Fold 3	0.9246	0.7673
Fold 4	0.8396	0.7557
Fold 5	0.8134	0.7443
Mean R^2 Score	0.8678	0.7982

Overall, Linear Regression is the better-performing model for this dataset.



CONCLUSION

The study found that Linear Regression is the most effective regression model for predicting y. Cross-validation and feature significance enhanced the study, emphasizing the relevance of model selection and exploration.