Stellar Classification

Milestone: Model Performance Evaluation and Interpretation

Group 16

Student 1 Sanidhya Karnik

Student 2 Digvijay Raut

617-407-1206 (Tel of Student 1)

857-492-3195 (Tel of Student 2)

karnik.san@northeastern.edu

raut.di@northeastern.edu

Percentage of Effort Contributed by Student 1: **50%**

Percentage of Effort Contributed by Student 2: **50%**

Signature of Student 1: **Sanidhya Karnik**

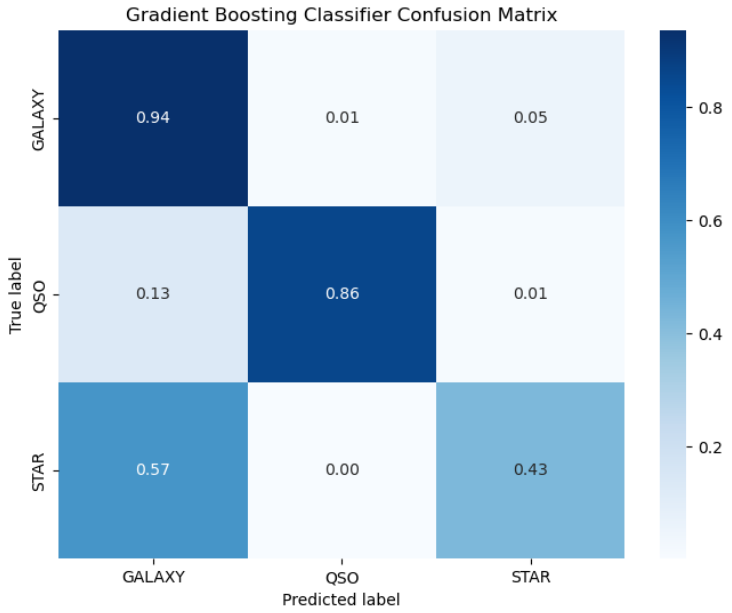
Signature of Student 2: **Digvijay Raut**

Submission Date: **03/15/2024**

In this report, we provide a comparative analysis of four distinct machine learning models: Gradient Boosting, Random Forest, Neural Network, and k-Nearest Neighbors (k-NN). These models were evaluated based on their performance in a classification task, using four key metrics: Accuracy, Precision, Recall, and F1 Score.

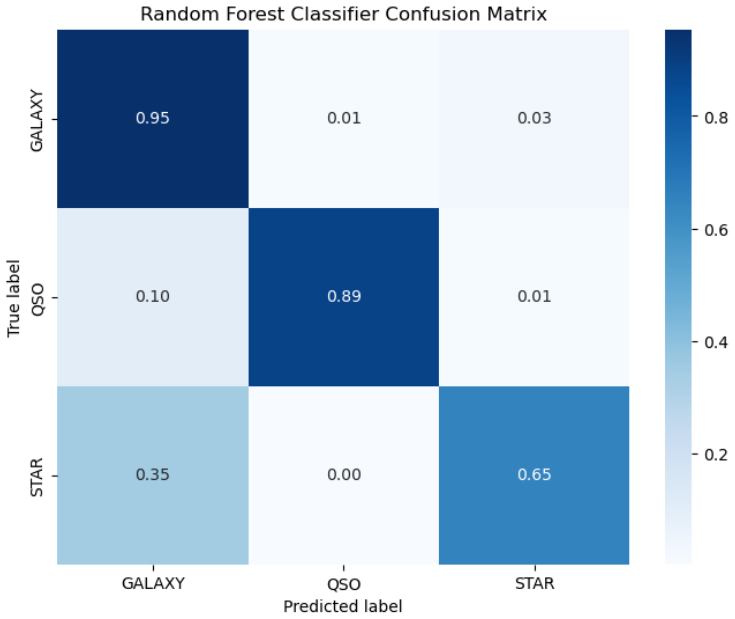
Gradient Boosting

Gradient Boosting exhibited an accuracy of 81.39%, precision of 81.17%, recall of 81.39%, and an F1 score of 80.04%. This ensemble learning method builds models sequentially, each correcting its predecessor, which is especially effective for handling bias and variance in data. However, its performance, while robust, was not the top among the models assessed, possibly due to its sensitivity to overfitting and the complex nature of the data. Confusion matrix for Gradient Boosting is as follows:



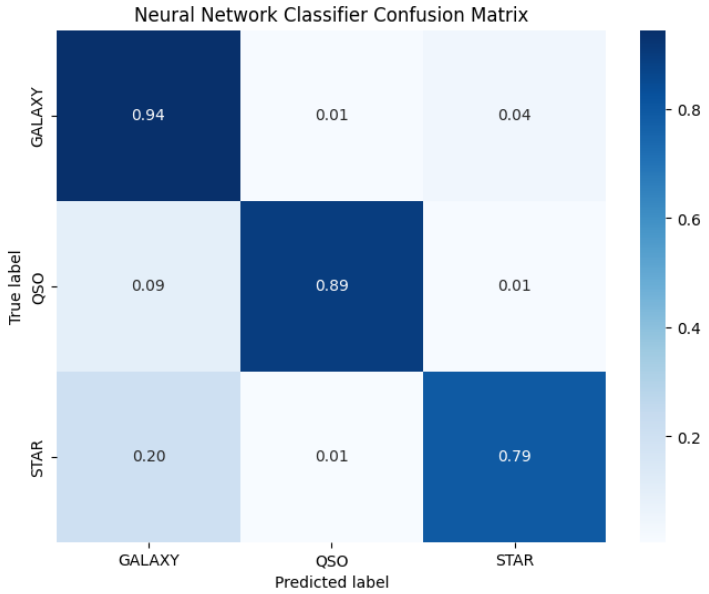
Random Forest

The Random Forest model showed improved performance with an accuracy of 87.86%, precision of 88.00%, recall of 87.86%, and an F1 score of 87.46%. As an ensemble of decision trees, Random Forest reduces overfitting risks and handles bias and variance more effectively than a single decision tree. Its higher scores across all metrics compared to Gradient Boosting suggest it is better suited for this classification task, likely due to its ability to manage complex interactions and dependencies in the data. Confusion matrix for Random Forest is as follows:



Neural Network

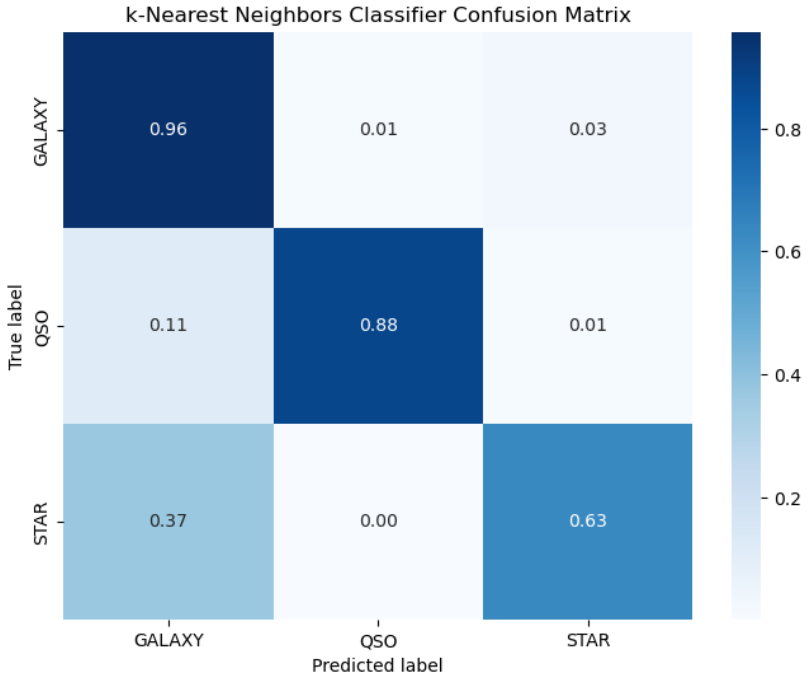
The Neural Network achieved the highest scores among the evaluated models, with an accuracy of 89.85%, precision of 89.89%, recall of 89.85%, and an F1 score of 89.68%. Neural networks are powerful tools for modeling complex relationships through their deep layers and many parameters. Their superior performance in this analysis underscores their capability in handling nonlinearities and interactions in data, albeit at the cost of requiring extensive computational resources and data for training. Confusion matrix for Neural Network is as follows:



k-Nearest Neighbors (k-NN)

k-Nearest Neighbors presented an accuracy of 87.27%, precision of 87.52%, recall of 87.27%, and an F1 score of 86.79%. k-NN is a simple, intuitive method that classifies samples based on

the majority vote of their neighbors. While it is less computationally intensive during training, its performance is slightly below that of the Random Forest and significantly lower than the Neural Network. This might be due to its dependency on a suitable distance metric and the challenge of choosing an optimal number of neighbors. Confusion matrix for k-NN is as follows:



Conclusion

The analysis demonstrates that Neural Networks offer the highest performance for this classification task, closely followed by Random Forest and k-Nearest Neighbors, with Gradient Boosting trailing slightly behind. Each method's effectiveness varies based on the specific characteristics of the data and the computational resources available. Therefore, the choice of model should consider the trade-off between accuracy and computational efficiency, alongside the specific requirements and constraints of the application.

Below plot contains area under curve for all four methods we used.

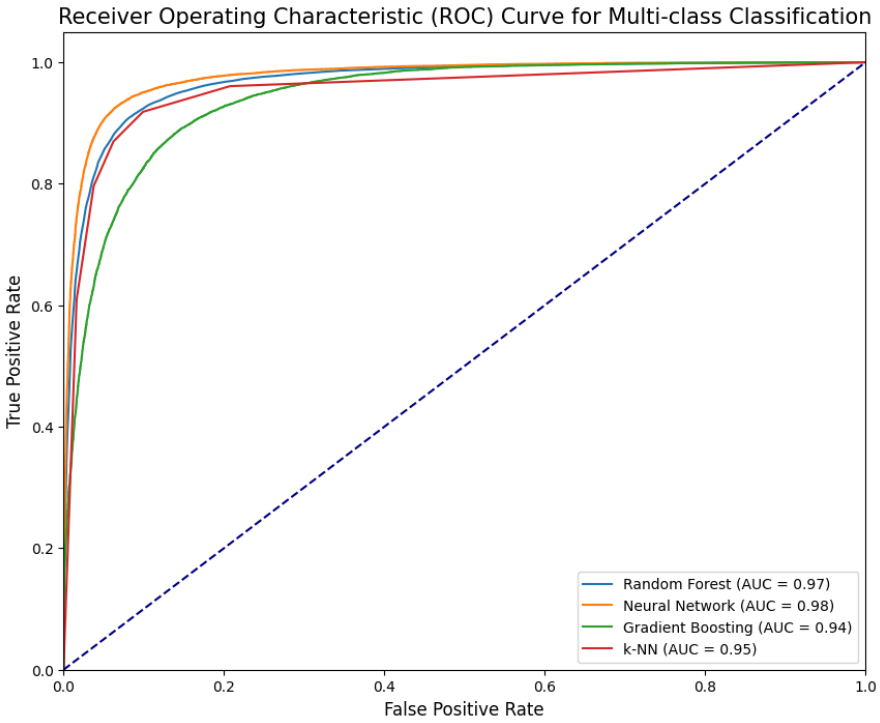


Table below shows accuracy, precision, recall and F1 score for all four models.

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.877000	0.878330	0.877000	0.872883
Neural Network	0.902556	0.902312	0.902556	0.901710
Gradient Boosting	0.813944	0.811650	0.813944	0.800372
k-NN	0.872667	0.875198	0.872667	0.867865