Performance Evaluation of Decision Tree Classifier Model and Support Vector Machine Model for the Classification of Moons Dataset

Parameter Tuning

Random Search & Grid search

Random search was used for parameter tuning in the SVM model and grid search was used for parameter tuning in the DT classifier model. Both parameter tuning methods select a best estimator which is a combination of the best parameter values for the model to use. The grid search method achieves this by creating a grid of parameters and testing them against one another to find the best parameters to use. The Random search finds the best parameters by randomly testing random parameter values with one another to find the best parameters.

Comparative Analysis

Mean Squared Error (MSE)

The mean squared error is a loss function that informs us of the average square difference between the model's predictions and the truth. A lower MSE indicates a higher level of accuracy in the model's predictions.

	SVM	DT
MSE	0.0125	0.0005

Here we can see that since the MSE is lower for the DT model that the DT model is out-performing the SVM model in this instance.

Accuracy

Accuracy measures the number of correct predictions that a model makes over the total number of predictions made by the model giving a percentage accuracy.

	SVM	DT
Accuracy	0.9875	0.9995

As can be seen in the accuracy scores above the DT model performs better for accuracy than the SVM model.

Precision

Precision is a measure of positive identifications, precision is calculated by taking true positives over true positives and false positives (TP/TP+FP).

	SVM	DT
Precision	0.9817425988965375	0.9994868287740628

The DT model performs better for the precision measure than the SVM model.

Recall

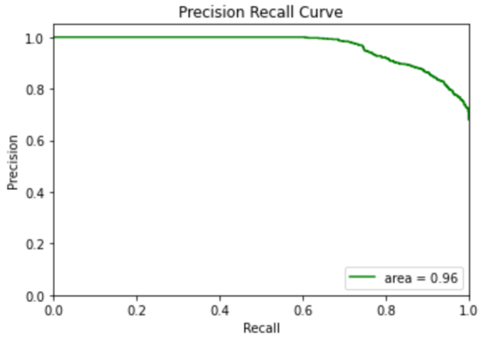
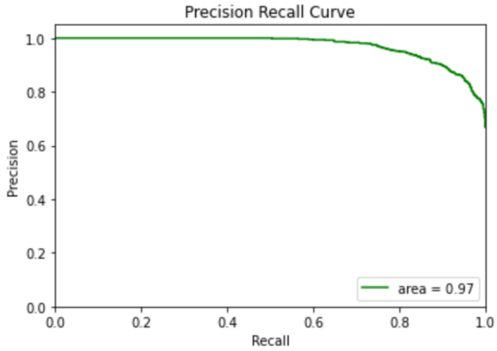
Recall is similar to precision but instead measures the proportion of true positives that were identified correctly (TP/TP+FN).

	SVM	DT
Recall	0.9870517928286853	0.9989868287740629

The DT model out-performs the SVM model for this measure.

Precision-Recall

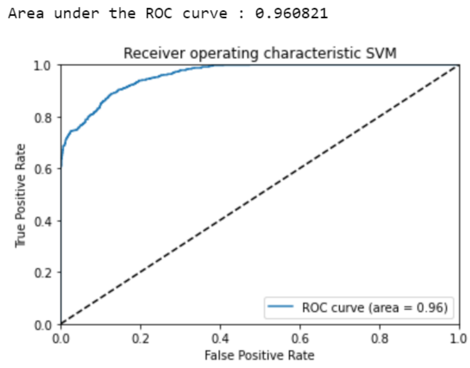
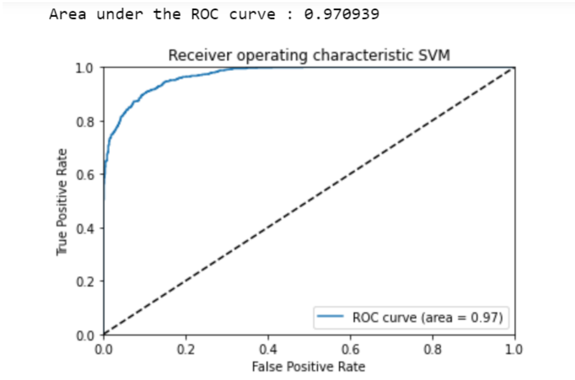
Below are the precision-recall curves for the SVM model and the DT model, which can also be seen in the jupyter notebook.

SVM	DT
	

The curve shows that the DT model performed better for both precision and recall than the SVM model.

Receiver Operating Characteristic Curve

The ROC curve displays the trade-off between the false positive rate and the true positive rate.
The ROC AUC is the probability that a randomly selected positive instance is ranked higher than a randomly selected negative instance. The closer that the ROC AUC is to 1 the better the performance of the model.
Below are the ROC curves for the SVM model and DT model.

SVM	DT
`Area under the ROC curve : 0.960821`	`Area under the ROC curve : 0.970939`

The DT model performed better on both the ROC curve and the ROC AUC than the SVM.

The ROC curve and the precision recall curve for the SVM model shows the same level of performance, as do the ROC and precision-recall curves for the DT model.

Conclusion

The DT classifier model performed better than the SVM model on all of the metrics, however with a mean difference between the models for all metrics of 0.012, it is arguable that the mean difference could be as a result of the datasets being randomly generated for each model and therefore not a constant between models. The difference could also be put down to the difference in parameter tuning for each model.

References

scikit-learn. n.d. *1.4. Support Vector Machines*. [online] Available at: <https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,Effective%20in%20high%20dimensional%20spaces.> .

Brownlee, J., 2021. *Tour of Evaluation Metrics for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>.

Chan, C., 2022. *What is a ROC Curve and How to Interpret It*. [online] Displayr. Available at: <https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/>.

Géron, A., n.d. *Hands-on machine learning with Scikit-Learn and TensorFlow*.

Self, G., 2019. *Understanding the 3 most common loss functions for Machine Learning Regression*. [online] Medium. Available at: <https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3#:~:text=The%20Mean%20Squared%20Error%20(MSE,out%20across%20the%20whole%20dataset.> .

Medium. n.d. *Support Vector Machine — Introduction to Machine Learning Algorithms*. [online] Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

GitHub. 2021. *handson-ml/05_support_vector_machines.ipynb at master · ageron/handson-ml*. [online] Available at: <https://github.com/ageron/handson-ml/blob/master/05_support_vector_machines.ipynb>].