## **Decision Trees**

**MILESTONES**

* Train-Test Split: Split data, train model, evaluate accuracy.
* 10-Fold Cross-Validation: Use cross-validation for model evaluation.
* Hyperparameter Tuning: Optimize model parameters using GridSearchCV
* Tree 1 (Best Params): Train with optimal parameters, evaluate test performance.
* Tree 2 (30% Test Data): Move 30% to test set, evaluate performance.
* Tree 3 (60% Test Data): Move 60% to test set, evaluate performance.
* Random Forests: Experiment with random forests, compare with decision trees.
* Overfitting Analysis: Compare accuracies, check for overfitting.

**FOR MAIN DATASET:**

* Accuracy: The accuracy of the model on the test set is 66.38%, which indicates that the model is able to correctly classify roughly two-thirds of the data. Cross-validation results show a similar performance (±0.0044).
* Precision: The precision is 0.6030, indicating that, out of all predictions the model made for a positive class, 60.30% were correct. Cross-validation precision is slightly better at 0.6062 ± 0.0030.
* Recall: The recall is 0.7383, meaning that the model successfully identified 73.83% of the true positive instances. Cross-validation recall is very close at 0.7359 ± 0.0049.
* F1 Score: The F1 score of 0.5781 reflects a balance between precision and recall, with a slightly higher recall than precision. The cross-validation F1 score is 0.5787 ± 0.0041, showing stability across different folds.
* ROC AUC: The ROC AUC of 0.8879 suggests that the model performs well in distinguishing between classes. The cross-validation ROC AUC is 0.8865 ± 0.0029, confirming consistency in the model's performance across different subsets of the training data

**Tree 1, Tree 2 and Tree 3**

**Tree 1: Performance Using Original Training and Test Set**

* **Test Accuracy**: 0.7675
* **Confusion Matrix**: The classifier demonstrates significant misclassifications in the second class (class 2), where the precision and recall are both zero, indicating that the model fails to predict this class effectively.
* **Classification Report**:
  + **Precision**: 0.81 (weighted average), indicating the model is good at making correct positive predictions for the majority classes.
  + **Recall**: 0.77 (weighted average), showing that the model captures a significant portion of the relevant instances for each class.
  + **F1 Score**: 0.76 (weighted average), reflecting a solid balance between precision and recall.
* **ROC AUC**: 0.8922, suggesting that the model has a strong ability to distinguish between the classes despite the misclassification issues for class 2.

**Tree 2: Performance with 30% of Data in Test Set**

* **Training Accuracy**: 0.7704
* **Test Accuracy**: 0.7672
* There is minimal performance degradation when using 30% of the data for testing, suggesting that the model generalizes well across different splits of the dataset.
* **Confusion Matrix**: The model still struggles with class 2, which appears with zero precision and recall. Misclassifications also increase for class 1.
* **Classification Report**:
  + **Precision**: 0.78 (weighted average), consistent with Tree 1, showing the model is reliable in its predictions.
  + **Recall**: 0.77 (weighted average), similar to Tree 1, indicating good identification of the correct instances.
  + **F1 Score**: 0.76 (weighted average), as in Tree 1, demonstrating a balanced performance.
* **ROC AUC**: 0.8904, slightly lower than Tree 1, but still showing strong discriminative ability.

**Tree 3: Performance with 60% of Data in Test Set**

* **Training Accuracy**: 0.7703
* **Test Accuracy**: 0.7689
* With 60% of the data in the test set, the model still performs relatively well, indicating its robustness even with a smaller training set.
* **Confusion Matrix**: Similar to Tree 2, the model continues to struggle with class 2, and misclassifications increase for other classes. The model seems to perform better for class 3, with a high recall (0.75).
* **Classification Report**:
  + **Precision**: 0.78 (weighted average), stable across trees, reflecting good performance in positive predictions.
  + **Recall**: 0.77 (weighted average), consistent, showing good generalization.
  + **F1 Score**: 0.76 (weighted average), maintaining balance in precision and recall.
* **ROC AUC**: 0.8893, slightly lower than Tree 2, but still a strong indicator of the model's ability to separate classes.

**GridSearchCV:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Class 1** | **Class 2** | **Class 3** | **Class 4** |
| Precision | 0.86 | 0.76 | 0.94 | 0.95 |
| Recall | 0.90 | 0.65 | 0.96 | 0.93 |
| F1-Score | 0.88 | 0.70 | 0.95 | 0.94 |
| Support | 15,729 | 3,505 | 52,808 | 47,525 |

**Overall Accuracy :** 93%

From the GridSearchCV results and the classification report, we can conclude the following:

1. Best Hyperparameters:
   * Using entropy as the criterion, max\_depth=None, min\_samples\_leaf=5, and min\_samples\_split=2 gave the best performance, balancing model complexity and accuracy.
2. Performance:
   * The best cross-validation accuracy was 0.8841, while the test set accuracy was 0.93, indicating strong generalization.
   * Precision and recall are high for most classes, but class 2 had lower recall (0.65), suggesting a class imbalance.
   * The F1-scores were good, especially for larger classes (class 3: 0.95, class 4: 0.94).
   * The ROC AUC of 0.9936 shows excellent model discrimination.
3. Key Insights:
   * A deeper tree (max depth=None) with controlled leaf size (min\_samples\_leaf=5) improved generalization.
   * Class 2 recall could be improved with techniques like class weighting or sampling.

**FOR SELECTED FEATURES DATASET:**

* Accuracy: The accuracy on the test set is 75.92%, which is a significant improvement over the previous result. The 10-fold cross-validation shows an average accuracy of 76.28%, with a small variance of ±0.0032.
* Precision: Precision is 0.7688, meaning 76.88% of predicted positive class instances are correctly classified. This is a strong result, and cross-validation shows a consistent precision of 0.7737 ± 0.0032.
* Recall: The recall is 0.7599, which indicates the model correctly identified 75.99% of true positive instances. Cross-validation recall is 0.7626 ± 0.0032, showing a stable performance across different folds.
* F1 Score: The F1 score of 0.7568 indicates a good balance between precision and recall. The cross-validation F1 score of 0.7602 ± 0.0033 is very close, indicating consistency across training subsets.
* ROC AUC: The ROC AUC of 0.8859 shows the model's ability to distinguish between the classes. Cross-validation ROC AUC is slightly higher at 0.8869 ± 0.0020, confirming the model’s robustness.

**Tree 1, Tree 2 and Tree 3**

**Tree 1: Performance Using Original Training and Test Set**

* **Test Accuracy**: 0.7592
* The classifier shows good accuracy and a balanced performance across different classes, especially with high recall for class 1 (0.91) and moderate precision for other classes.
* **Confusion Matrix**: The classifier correctly identifies the majority of instances from each class, with relatively fewer false positives and negatives.
* **Classification Report**:
  + **Precision**: 0.77 (weighted average), which indicates that, on average, the classifier is quite reliable.
  + **Recall**: 0.76 (weighted average), meaning the classifier is fairly consistent at identifying positive instances across classes.
  + **F1 Score**: 0.76 (weighted average), suggesting a balanced performance in terms of both precision and recall.
* **ROC AUC**: 0.8859, indicating that the model performs well at distinguishing between classes.

**Tree 2: Performance with 30% of Data in Test Set**

* **Training Accuracy**: 0.7624
* **Test Accuracy**: 0.7614
* There is a minor drop in performance when moving 30% of the data to the test set, suggesting the model's performance is relatively stable, even with a smaller training set.
* **Confusion Matrix**: The confusion matrix shows that the classifier's predictions are quite consistent with the classes, though there is a slight increase in misclassifications compared to Tree 1.
* **Classification Report**:
  + **Precision**: 0.77 (weighted average), similar to Tree 1, indicating stable reliability in prediction.
  + **Recall**: 0.76 (weighted average), comparable to Tree 1, showing consistent identification of classes.
  + **F1 Score**: 0.76 (weighted average), consistent with Tree 1, confirming a balanced model performance.
* **ROC AUC**: 0.8869, showing a slight improvement over Tree 1.

**Tree 3: Performance with 60% of Data in Test Set**

* **Training Accuracy**: 0.7634
* **Test Accuracy**: 0.7612
* The test accuracy drops slightly when moving 60% of the data to the test set, but it remains stable, showing that the model generalizes well even when a significant portion of data is held out for testing.
* **Confusion Matrix**: The model maintains its ability to predict most classes correctly, although misclassifications slightly increase compared to the smaller test sets.
* **Classification Report**:
  + **Precision**: 0.77 (weighted average), similar to the previous trees.
  + **Recall**: 0.76 (weighted average), still strong at identifying the right classes.
  + **F1 Score**: 0.76 (weighted average), maintaining a balanced performance.
* **ROC AUC**: 0.8867, very similar to Tree 2, showing consistent model performance in distinguishing classes

**GridSearchCV:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Class 1** | **Class 2** | **Class 3** | **Class 4** |
| Precision | 0.94 | 0.96 | 0.94 | 0.95 |
| Recall | 0.95 | 0.96 | 0.96 | 0.92 |
| F1-Score | 0.94 | 0.96 | 0.95 | 0.93 |
| Support | 52,808 | 52,808 | 52,808 | 52,808 |

**Overall Accuracy :** 95%

* The decision tree model with the selected hyperparameters performs excellently, with high accuracy (95%) on the test set and strong performance across all classes, as indicated by the precision, recall, and F1-scores.
* The ROC AUC of 0.9973 demonstrates that the model is very capable of distinguishing between the classes.
* The consistent high performance across metrics suggests that the chosen model configuration is effective for this dataset.

Best Hyperparameters:

Criterion: gini, max Depth: None, Min Sample Leaf: 5, Min Sample Split:2

**REFERENCES**

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>

<https://colab.research.google.com/drive/1a7ZfYvTSju41AXrEX5570-dFT8zGwVpH?usp=sharing>

<https://github.com/ageron/handson-ml2/blob/master/06_decision_trees.ipynb>

**NEXT STEPS**

* Understand about decision trees and do the tasks mentioned in lab task before the class