Checking the Proportion of Class Variable

In the churn_train dataset, the Exited variable indicates whether a customer has left the bank (1 for exited, 0 for retained). Here's a breakdown of the proportions:

- **0 (Not Exited)**: Approximately 79.49% of customers in the training dataset have not exited the bank.
- **1 (Exited)**: Approximately 20.51% of customers in the training dataset have exited the bank.

This suggests that the majority of customers are retained, while a smaller portion has left the bank.

Interpretation of Summary Statistics for Test Data

Examining the statistics for the Exited variable in the test dataset:

- **Min (Minimum)**: 0.0000
  - The smallest value (0) corresponds to customers who have not exited.
- **1st Qu. (First Quartile)**: 0.0000
  - 25% of the data points have a value of 0, indicating that at least 25% of customers have not exited.
- **Median**: 0.0000
  - The median (50th percentile) is 0, indicating that at least half of the customers in the test dataset have not exited the bank.
- **Mean**: 0.1756
  - The average (mean) value of the Exited variable is 0.1756, suggesting that approximately 17.56% of customers have exited the bank.
- **3rd Qu. (Third Quartile)**: 0.0000
  - 75% of the data points have a value of 0, indicating that up to the third quartile (75th percentile), the majority of customers have not exited.
- **Max (Maximum)**: 0.8244
  - There appears to be an anomaly here as the maximum value should typically be 1 in a binary dataset (0 or 1). This might require further investigation into the dataset for any discrepancies in data processing.

Interpreting the Decision Tree Structure for Training Data

- The decision tree begins with the feature IsActiveMember.
- If IsActiveMember is 1 (active member), the model predicts class 0 (not exited) with 4613 correct predictions out of 5266.
- If IsActiveMember is 0 (inactive member), the tree further splits on Age.
  - If Age is less than or equal to 44, the model predicts class 0 with 3482 correct predictions out of 4066.
  - If Age is greater than 44, the model predicts class 1 (exited) with 905 correct predictions out of 1201.

Evaluation on Training Data

- The decision tree comprises 3 nodes.
- There are 1533 errors out of 9000 cases, resulting in an error rate of 17.0%.

Confusion Matrix for Training Data

- **True Positives (TP)**: 609 (correctly classified as class 1)
- **True Negatives (TN)**: 6858 (correctly classified as class 0)
- **False Positives (FP)**: 296 (incorrectly classified as class 1)
- **False Negatives (FN)**: 1237 (incorrectly classified as class 0)

Attribute Usage

- **IsActiveMember** was utilized in 100% of the splits.
- **Age** was used in 48.74% of the splits.

Interpretation

1. **Tree Structure**: The decision tree is simplistic, consisting of only three nodes. The primary split is based on IsActiveMember, followed by Age for inactive members.
2. **Performance**: The model exhibits an error rate of 17%, indicating it accurately classifies 83% of the cases in the training set.
3. **Confusion Matrix**:
    - The model effectively predicts class 0 (not exited) with 6858 correct predictions but shows significant false negatives (1237).
    - Predicting class 1 (exited) is less accurate, with 609 correct predictions and 296 false positives.
4. **Attribute Importance**: IsActiveMember proves most critical, followed by Age.

This straightforward decision tree provides a clear framework for predicting customer churn based on whether a customer is an active member and their age. While its simplicity aids interpretability and reduces overfitting risk, enhancing performance might require additional features or model complexity to reduce false negatives.

Classification Tree Details

Non-Standard Options

- **Attempt to Group Attributes**: Suggests efforts to cluster similar features, potentially enhancing model simplicity and performance.
- **Minimum Number of Cases**: Each node in the decision tree required at least 400 samples, safeguarding against model complexity and overfitting.

Practical Implications

- **Simplicity and Interpretability**: A three-node tree facilitates understanding and communication of results.
- **Overfitting Prevention**: Attribute grouping and node size constraints mitigate overfitting, enhancing generalization to new data.

- **Model Accuracy**: While simplicity aids understanding, assessing validation or test datasets will confirm predictive efficacy in customer churn.

In conclusion, this classification tree model, leveraging 9000 samples and 10 predictors with a compact three-node structure, prioritizes simplicity and robustness. Attribute grouping and node size criteria curb overfitting, underscoring reliability in predicting customer churn pending validation assessment.

Interpreting the Decision Tree Structure for the Test Data

Confusion Matrix for the Test Data

1. **Total Observations in Table: 1000**
    - The dataset used for this evaluation contains 1,000 observations.
2. **Matrix Structure**:
    - **actual Exited**: The actual status of whether a customer exited or not.
    - **predicted Exited**: The predicted status from the model.

Confusion Matrix

actual Exited	predicted 0	predicted 1	Row Total
0 (Not Exited)	757	30	787
	0.962	0.038	0.787
	0.757	0.030	
---	---	---	---
1 (Exited)	151	62	213
	0.709	0.291	0.213
	0.151	0.062	
---	---	---	---
Column Total	908	92	1000

Interpretation for the Test Data

1. **Predicted Not Exited (0)**
    - **True Negatives (TN)**: 757
        - The model correctly predicted 757 customers who did not exit.
    - **False Negatives (FN)**: 151
        - The model incorrectly predicted 151 customers as not exited when they actually exited.
2. **Predicted Exited (1)**
    - **False Positives (FP)**: 30

        - The model incorrectly predicted 30 customers as exited when they did not.
    - **True Positives (TP)**: 62
        - The model correctly predicted 62 customers who exited.

Row Totals and Proportions

- **Row Totals**:
    - **Not Exited (0)**: 787 customers did not exit (actual).
        - **Proportion of correctly predicted non-exited**: $\frac{757}{787} \approx 0.962$ (96.2%)
        - **Proportion of incorrectly predicted non-exited**: $\frac{30}{787} \approx 0.038$ (3.8%)
    - **Exited (1)**: 213 customers exited (actual).
        - **Proportion of correctly predicted exited**: $\frac{62}{213} \approx 0.291$ (29.1%)
        - **Proportion of incorrectly predicted exited**: $\frac{151}{213} \approx 0.709$ (70.9%)

Column Totals and Proportions

- **Column Totals**:
    - **Predicted Not Exited (0)**: 908 customers predicted as not exited.
        - **Proportion of actual non-exited**: $\frac{757}{1000} = 0.757$ (75.7%)
    - **Predicted Exited (1)**: 92 customers predicted as exited.
        - **Proportion of actual exited**: $\frac{62}{1000} = 0.062$ (6.2%)

Model Performance Metrics for the Test Data

Based on the confusion matrix, we can calculate various performance metrics:

Accuracy: Measures how often the model's predictions are correct overall.

Precision: Measures how often the customers predicted to leave actually do leave.

Recall: Measures how well the model identifies actual leavers among all those who really left.

F1 Score: This is a balance between precision and recall. It’s the harmonic mean of precision and recall.

1. **Accuracy**:
   - $\text{Accuracy} = \frac{TP+TN}{Total} = \frac{757+62}{1000} = 0.819$ (81.9%)
2. **Precision** for Exited (1):
   - $\text{Precision} = \frac{TP}{TP+FP} = \frac{62}{62+30} = \frac{62}{92} \approx 0.674$ (67.4%)
3. **Recall** (Sensitivity) for Exited (1):
   - $\text{Recall} = \frac{TP}{TP+FN} = \frac{62}{62+151} = \frac{62}{213} \approx 0.291$ (29.1%)
4. **F1 Score** for Exited (1):
   - $\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.674 \times 0.291}{0.674 + 0.291} \approx 0.407$ (40.7%)

Conclusion

- The model's accuracy is relatively high at 81.9%.
- It performs well in predicting customers who do not exit (high true negative rate).
- However, the model struggles to correctly identify customers who exit (low recall and F1 score for the exited class).
- The low recall for the exited class suggests that the model misses many customers who leave, which could be critical for a churn prediction model aiming to retain customers. This indicates the need for further model tuning or considering other models/techniques to improve the prediction of customer churn.

Recommendation: Random Forest should be used to further evaluate