DigiPen Institute of Technology

CS 372: Decision Tree Assignment

Due: July 2nd, 2024

Objectives:

The objectives of this assignment are to help students:

- Learn how to implement Decision Tree classifier**.**
- Build Decision Tree using different types of features such continuous, ordinal, etc.
- Apply and compare various pre-pruning and post-pruning approaches.
- Apply Decision Tree classifier to real world problems.
- Evaluate their Decision Tree model.
- Analyze and interpret the result of Decision Tree model.

Problem description: Personal Loan classification problem

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget.

The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. The attributes can be divided accordingly:

- The variable **ID** does not add any interesting information. There is no association between a person's customer ID and loan, also it does not provide any general conclusion for future potential loan customers. We can neglect this information for our model prediction.

The binary features such as:

- Personal Loan - Did this customer accept the personal loan offered in the last campaign? **This is our target variable(The value that we would like to predict is the Personal Loan)**
- Securities Account - Does the customer have a securities account with the bank?

- CD Account - Does the customer have a certificate of deposit (CD) account with the bank?
- Online - Does the customer use internet banking facilities?
- Credit Card - Does the customer use a credit card issued by Universal Bank?

Continuous/ Interval features such as:

- Age - Age of the customer.
- Experience - Years of experience.
- Income - Annual income in dollars.
- CCAvg - Average credit card spending.
- Mortgage - Value of House Mortgage.

Ordinal features such as:

- Family - Family size of the customer

Categorical features such as:

- Education - education level of the customer; Undergrad; Graduate; Professional

Others are:

- ID
- Zip Code

Instructions

- Please, use Python for this assignment.
  - You need to implement Decision Tree using sklearn.
  - You could use the following libraries: sklearn.tree, sklearn.model_selection , Matplotlib, NumPy, Pandas, graphviz, Seaborn, SciPy.

Question 1

❖ Download the ***Bank Personal Loan Information*** data from course Moodle.
❖ Load Data from CSV file to Pandas data frame.
❖ To implement your Decision tree model, you need to:
  ➢ Identify your target feature and descriptive features. In our case, Personal Loan will be our target feature and the rest of the features (descriptive features) will be used to predict Personal loan.

Data Exploratory Analysis [15 points].

- **[6 points]** Check for:
  i. **[2 point]** duplicates, show the last duplicates. Choose one of the methods we studied in the class to deal with duplicates records (if any) and explain how and why you did that.
  ii. **[2 point]** missing data in the data frame. Choose one of the methods we studied in the class to deal with missing values in Decision Trees, if any. Make sure to explain what you used to fill out missing data. Note that zero is a value not missing data.
  iii. **[2 point]** Check for Outliers using Boxplot graphs. Make sure to add necessary labels for each plot. Upon detection, show and explain how you dealt with outliers, justify your choice.
- **[4 points]** Draw scatter plot matrix for all features and comment on the plot.
- **[5 points]** Heatmap correlation: Draw a heatmap plot between all independent variables and dependent variables. Comment on your heatmap (i.e. the strength and the direction of the correlation between features).

Decision Trees [33 points]

- **[10 points]** To use categorical data with scikit learn in Decision Trees, we have to use One-Hot Encoding trick to convert categorical columns into multiple columns of binary values. There are two popular methods that you could use for this purpose, namely:
  - ColumnTransformer() (from scikit-learn)
  - get_dummies() (from pandas)

  Note that One-Hot Encoding converts a column with more than 2 categories, and hence, we don't need to process categorical features with only two categories.

- **[3 points]** Split your dataset 80/20. Train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction. Make sure that if I run the algorithm again, I will get the same split and also the proportion of the classes are preserved in the split. Shuffle the data before split.

- **[10 points]** Build and Compare two Decision Trees, based on:
  - **[5 points] Gini Impurity.**
  - **[5 points] Entropy.**

- **[10 points]** Visualize the fully grown Decision Trees you built (tree graph).(5 points for each tree)

Model Evaluation [10 points]

- **[6 points] To evaluate the performance of your classifier use:**
  - ✓ **[2 points]** execution time.
  - ✓ **[2 points]** Accuracy.
  - ✓ **[2 points]** confusion matrix.
- **[2 points]** Analyze, explain, and comment on your result.
- **[2 points]** How does the Number of training sample affect performance (accuracy, time, etc.)? Explain and draw a graph.

❖ **Tree Pruning [37 points]**

- Implement the following pruning methods for both Decision Trees (Gini and Entropy):
  - **[5 points]** maximum depth
    - What is the best depth limit to use for this data? Use 5-fold cross validation for selection.
  - **[5 points]** minimum number of samples. Use 5-fold cross validation to decide.
  - **[5 points]** cost complexity pruning
- **[10 points]** Compare the overall accuracy of Decision Trees using different pruning methods in terms of accuracy and time. Next, make sure to compare that with fully grown tree.
- **[12 points]** Plot and interpret all pruned trees.

Notes

- You are working with imbalance data.
- You might need to ignore some features by studying the importance of different features in classification results.

Submit the following:

Final Submission (Due July 2nd, 2024):

Your notebook should include two types of cells: Code cells and markdown cells. Use the first to write code and comments that explain what the code is doing. And the later can be used to add text for analysis, summary, and conclusion.

a. **Code.**
   - Name your file: Assignment2_GroupName.ipynb. such as:
     - Assignment2_Group1.ipynb.
   - **[2 points]** Your code should contain:
     - appropriate comments to facilitate understanding.

    - Clear execution Instruction.
    - Write the name of the student who wrote each piece of code.

b. [3 points] Summary at the end:

Make sure to include a full explanation of the results and analysis. Moreover, include:

- A high-level description of how your code works.
- The evaluation matrices you used such as accuracies you obtain under various settings.
- Discussion:
    - Explain which options work well and why.
    - If all your evaluation matrices are low, tell us what you have
    - tried to improve them and what you suspect is failing,
    - challenges, and explain your output.
- Conclusion and summary