Exercise 5: Classification Tree

Workflow

1. Download the .ipynb files and data files posted corresponding to this exercise and store them in a single folder.
2. Open and explore the .ipynb files (notebooks) that you just downloaded and go through "Preparation" as follows.
3. The walk-through videos posted on NTU Learn (under Course Content) may help you with this "Preparation" too.
4. Create a new Jupyter Notebook, name it MatID_Exercise5_solution.ipynb, where "MatID" is your Matric Number.
5. Solve the "Problems" posted below by writing code and comments in MatID_Exercise5_solution.ipynb notebook.
6. Submit the Notebook MatID_Exercise5_solution.ipynb to your respective Lab Group's Course Site on NTU Learn. 7. Talk to your TA at the Lab Session regarding submission portal and/or procedure before you submit your solution.

Try to solve the problems on your own. Take help and hints from the "Preparation" codes and the walk-through videos. **If you are still stuck, talk to your TA in the Lab Session to get help/hints.** Try not to discuss this with your classmates.

Note : Don't forget to import the Essential Python Libraries required for solving the Exercise. Write code in the usual "Code" cells, and notes/comments in "Markdown" cells of the Notebook. Check the preparation notebooks for guidance.

Preparation

M4 ClassificationTree.ipynb Check how to perform basic Classification on the Pokemon data (pokemonData.csv)

Objective

Note that our Housing Data has a Binary (two-level) Categorical Variable named "CentralAir", with values "Y" and "N". In the previous sessions, we have seen some numeric variables in this dataset that are important to predict "SalePrice". In this Example Class, we will try to predict if a house has Central Air Conditioning or not using some other variables. This assignment is **NOT** graded.

Problems

Download the dataset **train.csv** and the associated text file **data_description.txt** posted with this Exercise.

Problem 1: Predicting CentralAir using SalePrice

Import the complete dataset "train.csv" in Jupyter : `houseData = pd.read_csv('train.csv')` Use the following variables from the dataset in this problem : `SalePrice` and `CentralAir`

a) Plot the distribution of `CentralAir` to check the imbalance of Y against N. Print the ratio of the classes Y : N.

b) Plot `CentralAir` against `SalePrice` using `any appropriate bivariate plot` to note the mutual relationship.

c) Import Classification Tree model from Scikit-Learn: `from sklearn.tree import DecisionTreeClassifier`

d) Partition the dataset `houseData` into two "random" portions: Train Data (1100 rows) and Test Data (360 rows).

e) Training: Fit a Decision Tree model on the Train Dataset to predict the class (Y/N) of CentralAir using SalePrice.

f) Visualize the Decision Tree model using the `plot_tree` function: `from sklearn.tree import plot_tree`

g) Predict `CentralAir` for the train dataset using the Decision Tree model and plot the Two-Way Confusion Matrix.

h) Print accuracy measures of the Decision Tree model, including its Classification Accuracy, True Positive Rate, True Negative Rate, False Positive Rate and False Negative Rate, based on the confusion matrix on train data.

i) Predict `CentralAir` for the test dataset using the Decision Tree model and plot the Two-Way Confusion Matrix.

j) Print accuracy measures of the Decision Tree model, including its Classification Accuracy, True Positive Rate, True Negative Rate, False Positive Rate and False Negative Rate, based on the confusion matrix on test data.

Problem 2: Predicting CentralAir using Other Variables

Perform all the above steps on "CentralAir" against each of the variables "GrLivArea", "OverallQual", "YearBuilt", one-by-one to perform individual Binary Classifications and obtain individual univariate Decision Tree Models in each case. Consider all predictor variables "GrLivArea", "OverallQual", "YearBuilt" as *Numeric* in case of this classification problem.

Problem 3: Best Uni-Variate Model to Predict CentralAir

Compare and contrast the four models in terms of Classification Accuracy, True Positive Rate and False Positive Rate on both Train and Test Data to comment on which univariate classification tree you think is the best to predict "CentralAir".

Feel free to comment throughout the notebook (using markdown) to explain and justify your solution and conclusion.

Extra Resources

You may read more about the `DecisionTreeClassifier` model you use in this exercise in the following references.

`DecisionTree` : https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Other Tree Models (Scikit Learn) : https://scikit-learn.org/stable/modules/classes.html#module-sklearn.tree

Bonus Problems

1. Note that the DecisionTreeClassifier() model can take more than one Predictor to model the Response variable. Try to fit a Decision Tree model to predict "CentralAir" using all the four variables "SalePrice", "GrLivArea", "OverallQual" and "YearBuilt". Print the Classification Accuracy of this multi-variate model on Train and Test datasets, and check the model's reliability of prediction on Train and Test data using the confusion matrices.

2. Fit a Decision Tree model to predict "CentralAir" using all numeric variables in the given dataset. You may use all numeric variables from Exercise 2. Print the Classification Accuracy of this multi-variate model on Train and Test data, and check the model's reliability of prediction on Train and Test data using the confusion matrices.

3. Are False Positive Rates of the various Decision Tree models significantly higher/lower than False Negative Rates? Does this have anything to do with the unbalanced classes (Y : N) of 'CentralAir'? Experiment and think about it.