AccelerateAI

Data Science Global Bootcamp

Assignment 09

Decision Trees

Q1. Decision Tree Classification – Credit Data

Consider the dataset in 01_credit_data.csv file with the following:

- A1 - A15 = Attributes
- T - Target (positive or negative credit)

Split the data into train and test (75% - 25%) and train a decision tree using sklearn's DecisionTreeClassifier() with 2 different methods:
1) ID3 (criterion = 'entropy')
2) CART (criterion = 'gini')

Find the accuracy on test data for the 2 different methods. Why do you think they are so?

Q2. Decision Tree Regression – MBA Starting Salary

The average starting salary of MBA graduates in US is provided in the file 02_MBA_Starting_Salary.xlsx.

Use the attributes below to predict the Avg Starting salary:
Type
Enrollment
Avg GMAT
Resident Tuition, Fees
Pct International
Pct Female
Pct Asian American
Pct Minority
Pct with job offers

1) Train a Decision tree by dropping the rows with missing values
2) Impute the missing values in each column using KNN imputer, and then train a model
3) Compare the score of both the above models. Use 5-fold CV score in both cases, with same model hyper-parameters (depth, etc)

Q3. Ensemble Modeling – Mobile Price

We want to predict the price of mobile phone(range) based on the characteristics of the phone like memory, battery power, camera specification etc. The data for about 2000 phones is provided in 03_mobile_price.csv

1) Train a Random Forest classifier to predict the mobile price. What is the accuracy you get with a 5fold CV on the Dataset.
2) Now, train a gradient boosting classifier using the same data. How does the score compare with random forest model?

Note that this is multi-class classification.

The data files can be found here: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/tree/main/ClassAssignment/Assignment09