**JEREMY PETERSON DUTOYA**

**PR. MENG-SHEN HSIEH**

**PMBA-8358-OLA: DATA-DRIVEN STRATEGIES FOR BUSINESS**

**CART HOMEWORK ASSIGNMENT**

**PART A: USE CART TO FIT A REGRESSION TREE**

Objective: use the WineTrain.csv to build a regression tree and then use the tree to predict wine prices from the test data Wine Test.csv (Note: WineTrain.csv contains data from 1952 to 1978 and WineTest.csv contains data from 1979 to 1980. You are using the in-sample data from 1952 to 1978 to fit the tree and then make the out-of-sample predictions from 1979 to 1980).

1. Use the following variables as the input variables to build a regression tree to predict wine price:
   1. *WineRain*
   2. *AGST*
   3. *HarvestRain*
   4. *Age*

Set the model tuning parameters: Min Split = 0, Max Depth=30, Min Bucket=1, Complexity=0.01

Diagram

Description automatically generated

1. Based on the rules of splitting the tree, state the rule which leads to the **highest price**?

According to the Decision Tree created and based on the rules of splitting the tree, the rule which leads to the highest price (i.e., the extreme right node with an average price of 8,5) is:

**AGST >= 17, Age >=21 and Winter Rain >= 760**

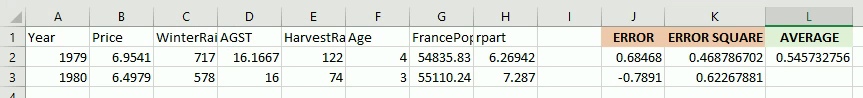
In addition, and as I paid very close attention to you explanation in the videos provided in this module, I navigated the R Data Miner “rules” tab to have a more precise and better idea of the different rules and I have found that the following leads to the highest price:

A picture containing letter

Description automatically generated

1. Predict the wine price using the **test** data.

Define **Error** **= Actual Price – Predicted Price.** What is the average of the squared errors?



As showcased in the above picture the average of the squared errors is approximately 0.55.

**PART B: USE CART TO FIT A CLASSIFICATION TREE**

Objective: use Wikipedia.csv to build a classification to determine if the edit of a Wikipedia page was vandalism.

Definitions of the variables of the dataset:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Vandal | 1 if this edit was vandalism, and 0 if not |
| Minor | 1 if the user marked this edit as a minor edit and 0 if not |
| LoggedIn | 1 if the user made this edit while using a Wikipedia account, and 0 if they did not |
| HTTP | 1 if the edit contained a web address, and 0 otherwise |
| NumWordsAdded | The number of unique words added |
| NumWordsRemoved | The number of unique words removed |

**Please implement the following tasks:**

Partition the data into train and test dataset with 70% and 30% of the entire data respectively.

Set the model tuning parameters:

**Min Split = 0, Max Depth=30, Min Bucket=7, Complexity=0.001**

The predicted outcome of each node is determined by the majority rule in that node in which there are at least 7 observations.

1. Build a CART model to predict **Vandal**, using all of the other variables as the independent variables. You should use the training set to build the model and use the test set for predictions.
2. Plot the classification tree.
3. Which variables were used in the tree? (what variables are used in the rules for splitting the branches of the tree?)

Diagram

Description automatically generated

The variables used in the rules for splitting the branches of the tree are LoggedIn, NumWordsAdded (x2), NumWordsRemoved, HTTP, Minor as you can see below with Rule #91

Text

Description automatically generated with medium confidence

1. Use the test data to predict the vandalism. What is the model accuracy in terms of percentage of accuracy?

Graphical user interface, text, application, email

Description automatically generated

As you explained, the diagonal entries represent the correct predictions and the off-diagonal entries represent the incorrect predictions, so in terms of percentage of accuracy the model has a 74% of accuracy.

*We have 2 ways of calculating this: 48.6 + 25.4=74% with the diagonal or 100-26% overall error=74%.*