# Lab Decision Tree and Ensembles

Please submit your answers in this word document, including detailed discussion for each question with supportive screenshots of your analysis in software.

You have been provided with the Cleaned Marketing from Database file from your client, a local nonprofit organization. This file contains information on 9,686 customers with demographic information such as age, gender, if they are a homeowner, home value, income, veteran status, as well as information on their interactions with the organization such as if they are donated last year (Target B), donation amount last year (Target D), average gift over the last 36 months, their most recent gift, and so forth. Check that all Variables are set to Interval, except for DemGender, DemHomeOwner, ID, StatusCat96NK, and TargetB, which are set to Nominal. The Validation column is the indicator of the validation observations. Refer to the data dictionary provided with the assignment for details.

Your clients have requested for you to develop a decision tree model to identify who they should send a soliciation to this year using Target B (if the customer donated last year). Target D should NOT be used in any model.

The average donation received last year was $15.62 and it costs $0.68 to send each mailing.

* Create a new SAS EM diagram “Decision Trees Homework”. Import the Cleaned Marketing from Database file and rename the import node “Marketing”. A SAS code node can be connected to the file import node to partition data using the Validation variable and following SAS code.

DATA&EM\_EXPORT\_TRAIN &EM\_EXPORT\_VALIDATE;

SET&EM\_IMPORT\_DATA;

IF Validation=1 Then output&EM\_EXPORT\_VALIDATE;

else if Validation=0 Then output&EM\_EXPORT\_TRAIN;

Run;

* Prepare 4 models
  + Decision tree with 2 Branches and 6 Depth
  + Decision tree with 4 Branches and 12 Depth
  + Gradient Boosting – default settings
  + Random Forest – default settings
* Compare the 4 models by Misclassification Rate of the validation data

1. Please provide a screen shot of your Decision Tree Homework Diagram. Make sure that the project name (which includes your name) is visible.

A screenshot of a computer

Description automatically generated

1. Which model was selected? Provide a screen shot of the output. Make sure that the target variable is Target B.

Based on the validation data misclassification rate, I got Decision Tree with 4 branch and 12 depth best model out of all other models with misclassification rate of 0.41347 when predicting TargetB.

A screenshot of a computer

Description automatically generated



1. Provide the plot of the model change in Misclassification Rate vs. Number of Leaves plot (Subtree Assessment Plot) for the best of the two decision tree models (based on their performance on the validation data). Remember to include your first and last name in the graph.

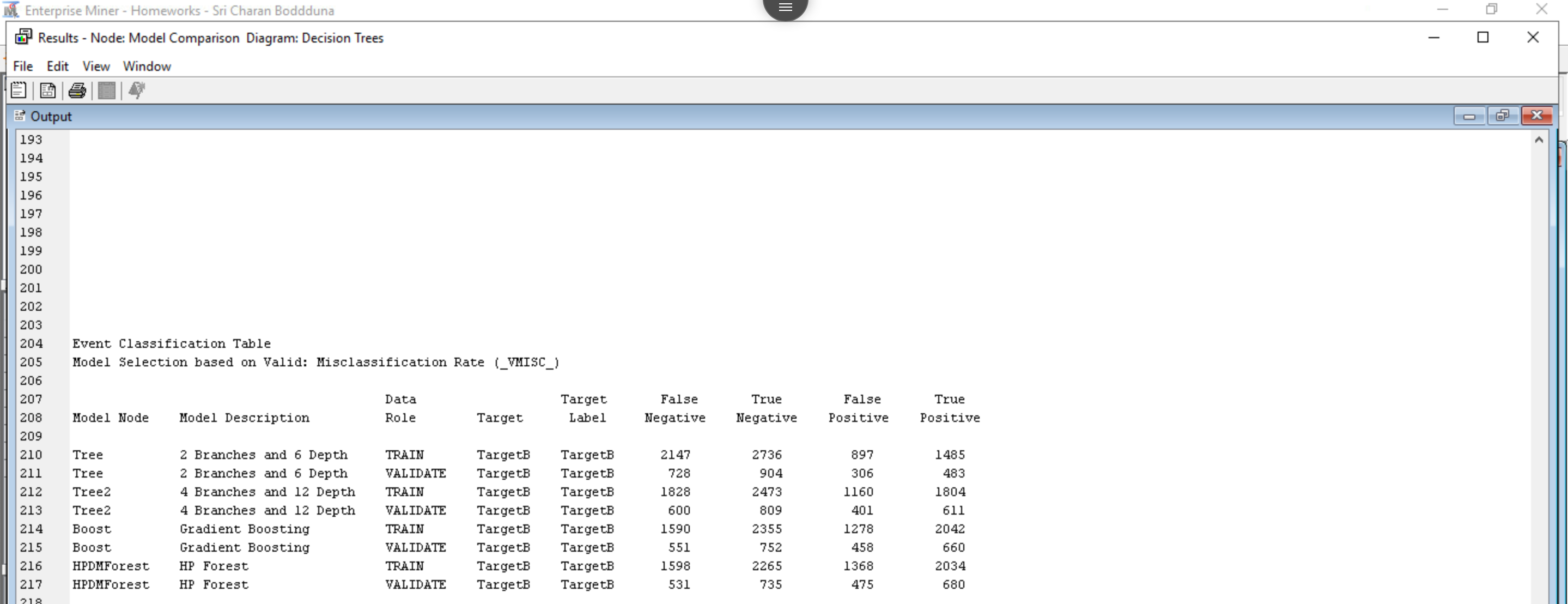
The best model between two decision tree models is decision tree with 4 branch 12 depth. For the same model we can see the plot in the below screenshot. We can see validation misclassification rate (MR) follows training MR till four leaves in the model. Later, valid MR leads training MR which means the model is performing better and produces good prediction till 20 leaves. After that, the model gets into overfitting as splits in the tree are more and the gap between valid MR and train MR is high.

A graph on a computer screen

Description automatically generated

1. Complete the following table for validation data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FN | TN | FP | TP | Misclassification Rate |
| HP Forest | 531 | 735 | 475 | 680 | 0.41553 |
| Gradient Boosting | 551 | 752 | 458 | 660 | 0.41677 |
| 4 Branch 12 Depth | 600 | 809 | 401 | 611 | **0.41347 (best)** |
| 2 Branch 6 Depth | 728 | 904 | 306 | 483 | 0.42710 |



1. Calculate the profit for each of the models ignoring opportunity costs on the validation data. Show your calculations to support your answer.

**Model: 2 Branch 6 Depth**:

TP = 483; FP = 306

average donation received =$15.62

costs = $0.68.

donation received \* TP + costs \* FP

483 \* 15.62 + 306 \* 0.68 = **$ 7752.54**

**Model: 4 Branch 12 Depth**

TP = 611; FP = 401

average donation received =$15.62

costs = $0.68.

donation received \* TP + costs \* FP

611 \* 15.62 + 401 \* 0.68 = **$ 9816.5**

**Model: Gradient Boosting**

TP = 660; FP = 458

average donation received =$15.62

costs = $0.68.

donation received \* TP + costs \* FP

660 \* 15.62 + 458 \* 0.68 = **$ 10620.64**

**Model: Random Forest (HP Forest)**

TP = 680; FP = 475

average donation received =$15.62

costs = $0.68.

donation received \* TP + costs \* FP

680 \* 15.62 + 475 \* 0.68 = **$ 10944.6**

1. Is the model with the largest profit the same model as the one selected with the best misclassification rate? Why do you think this is the case?

No, the best classification rate is found to be for decision tree with 4 branch and 12 depth and the largest profit is for random/HP forest model.

I think random forest forms multiple decision trees of different branches and depths and sums the results and applies the majority rule. Few features can be missed out by going into incorrect trees which may lead to increased false positives in the model, thereby increased misclassification rate.

On the other hand, we can see decision tree with 4 branches and 12 depths have least misclassification rate since we limited the number of branches and depths which can act as stopping rule this leads to better false negatives.

1. Provide a screen shot of the Variable Importance for the Gradient Boosting model.

A screenshot of a computer

Description automatically generated

1. Provide a screen shot of the Variable Importance for the 2 Branch 6 Depth model.

A screenshot of a computer

Description automatically generated

1. Provide a screen shot of the Variable Importance for the 4 Branch 12 Depth model.

A screenshot of a computer

Description automatically generated

1. Provide a screen shot of the Variable Importance for the Random Forest model.

A screenshot of a computer

Description automatically generated

1. Looking at the five most important variables of each model on the training data, what do you think would be the two most important variables across all models? Focus on the ‘Importance’ or ‘Gini Reduction’ on Training data (Not the Validation Gini Reduction or Validation Importance).

|  |  |  |  |
| --- | --- | --- | --- |
| **2 branch 6 depth** | **4 branch 12 depth** | **Gradient Boosting** | **Random Forest** |
| GiftCnt36 | GiftCnt36 | DemMedHomeValue | GiftCnt36 |
| GiftTimeLast | DemMedHomeValue | GiftTimeLast | GiftAvgCard36 |
| GiftCntCardAll | GiftTimeLast | GiftCnt36 | GiftTimeLast |
| PromCnt12 | PromCntCard12 | GiftAvgCard36 | GiftAvgLast |
| StatusCatStarAll | GiftCntCardAll | GiftAvgLast | DemMedHomeValue |

Two most important variables across all models are:

1. **GiftCnt36**
2. **GiftTimeLast**