Assignment 2: Decision Trees (100 points)

INFO3237: Business Analytics II

Instructor: Dr. Xue Guo

**Purpose:** To build and evaluate decision trees

**Description:** Using data from 2010 Congressional elections, we intend to build a classifier that would predict the election’s outcome. The data set includes information about the campaign funds, social media (Twitter, Facebook, and YouTube) campaigns, and demographics (age, gender) of 941 candidates who were in race in the general elections for The 112th House of Representatives seats.[[1]](#footnote-0)

**Instructions:** You need to follow these steps:

1. In Canvas, navigate to Assignments and then Assignment2
2. Download and save the data set election\_campaign\_data.csv
3. Read the file: data <- read.csv("election\_campaign\_data.csv", sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))
4. Drop the following variables from the data: "cand\_id", "last\_name", "first\_name", "twitterbirth", "facebookdate", "facebookjan", "youtubebirth".
5. Convert the following variables into factor variables: “twitter”, “facebook”, “youtube”, “cand\_ici”, and “gen\_election”.
6. Bear in mind that “twitter” equals 1 if the candidate had a Twitter campaign during the election and zero otherwise. The same would apply for “facebook” and “youtube”. “opp\_fund” is the total campaign fund of the opposing candidate. “gen\_election” is our target variable which takes value of “L” when the candidate lost the election and “W” when the candidate won the election. “ttl\_receipts” represents the total receipts the candidate received and “cand\_ici” represents the Incumbent challenger status (C = Challenger, I = Incumbent, O = Open Seat)
7. Remove all of the observations with any missing values.
8. Answer the following questions:
   1. **(4 points)** How many attributes this data set has?

There are 21 attributes

* 1. **(4 points)** How many records this data set has?

There are 929 records

* 1. **(4 points)** How many classes the class attribute (gen\_election) has?

There are 2 classes

* 1. **(4 points)** How many candidates had a Facebook campaign account (column facebook=1)?

There are 242 candidates

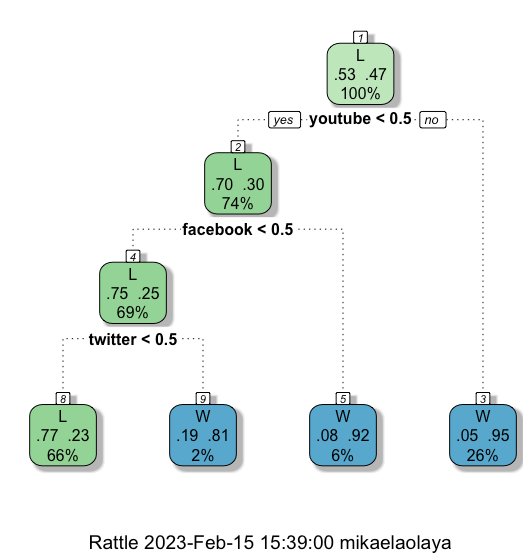
* 1. **(4 points)** How many candidates who had a twitter campaign account won the election?

There are 185 candidates that had a twitter account and won the election.

* 1. **(4 points)** How many incumbent candidates won the election?

339 incumbents won the election

1. Use “rpart” and “rpart.plot” packages to create decision trees and answer the questions below **(please set.seed = 100).**
   1. **(8 points)** Create a decision tree using variable gen\_election as the target (class) variable and the variables twitter, facebook, and youtube as the predictors. Paste the tree in the space below.



* 1. **(4 points)** Use “summary” to check which variable had the highest importance in the decision tree. List the variables based on the highest to lowest importance.

Youtube, Facebook, and Twitter

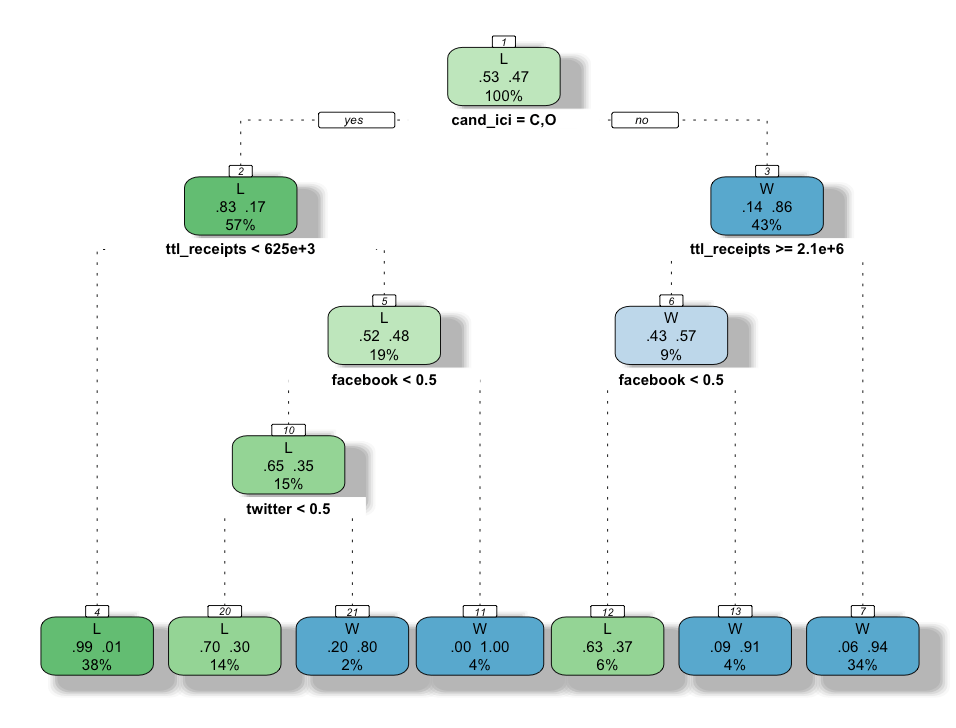
* 1. **(8 points)** How many of the candidates were correctly classified in the decision tree?

538 candidates were correctly classified

* 1. **(4 points)** Which leaf/terminal node is the purest node in the decision tree? Use GINI Index.

Node 3

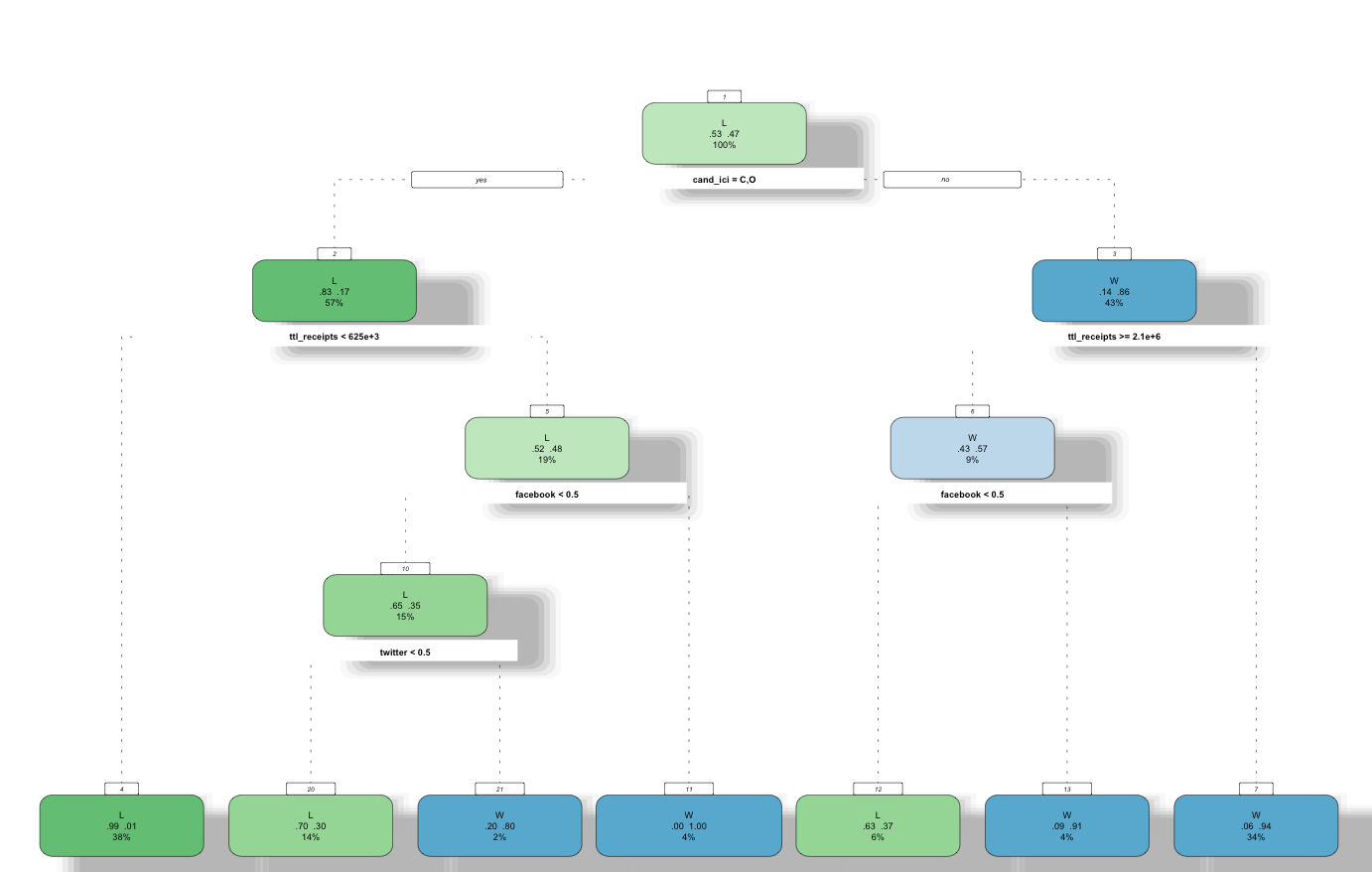
1. Now add ttl\_receipts and cand\_ici to the list of predictors.
   1. **(8 points)** Create and paste the decision tree in the space below.



* 1. **(4 points)** Use “summary” to check which variable had the highest importance in the decision tree. List the variables based on the highest to lowest importance.

cand\_ici, ttl\_receipts, youtube, facebook, and twitter

* 1. **(6 points)** Find the min error tree. Please report the cross validation set error and paste the tree in the space below.



The cross validation set error is the smallest xerror which would be 0.24590

1. Use packages “caret” and “e1071” to compare the two decision trees (the first one from 9.1 and the second one from 10.3) that you’ve created to answer the following questions. Use threshold <- 0.5. (Notes: Please use **set.seed = 100** for all the following answers, otherwise you will lose 10 points)
   1. **(4 points)** What is the accuracy of the first decision tree?

0.8309

* 1. **(8 points)** Fill out the confusion matrix below using the first decision tree:

|  | Reference | | |
| --- | --- | --- | --- |
| Predicted |  | 0 | 1 |
| 0 | **140** | **39** |
| 1 | **8** | **91** |

* 1. **(4 points)** What is the accuracy of the second decision tree?

0.8885

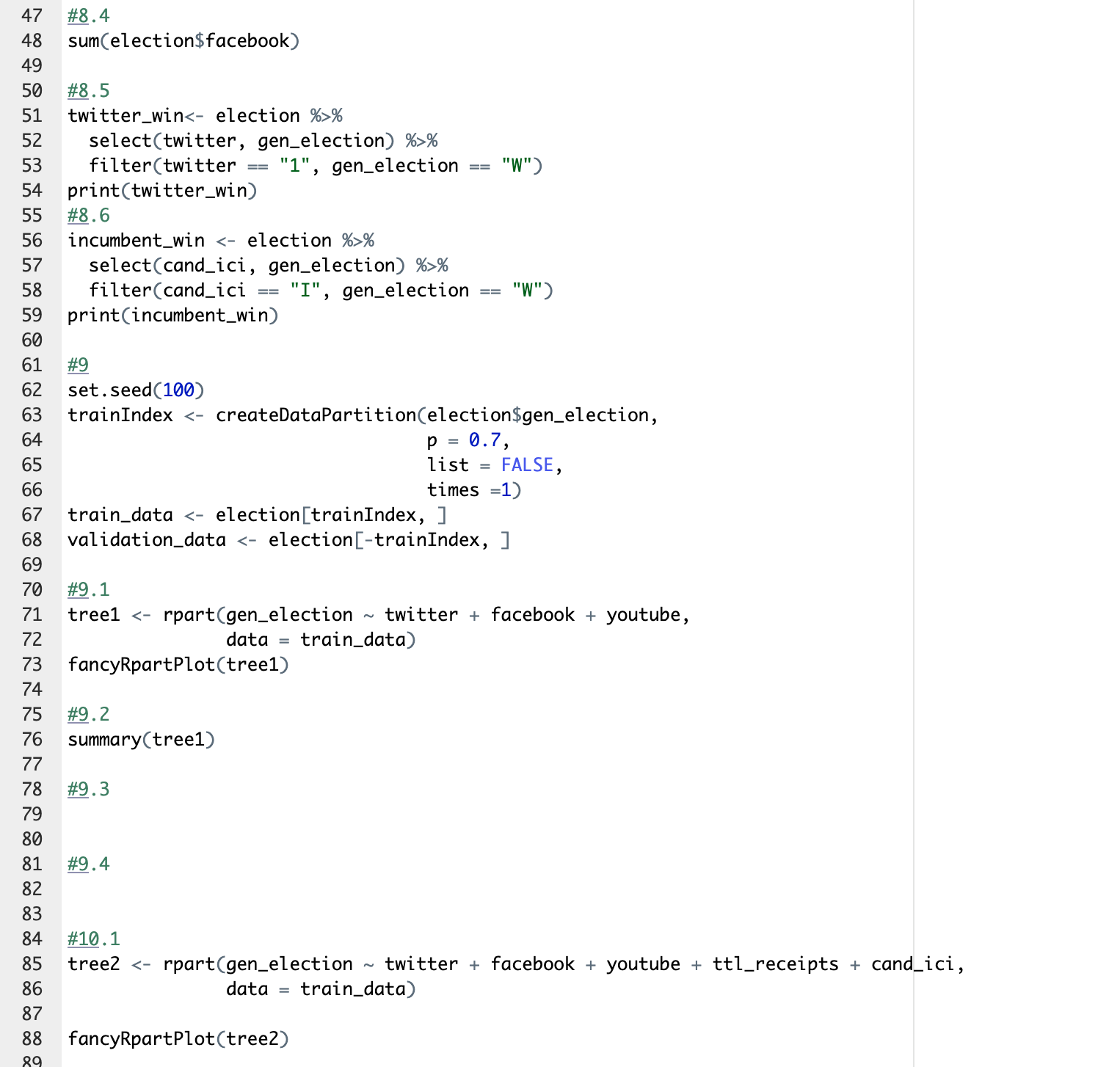
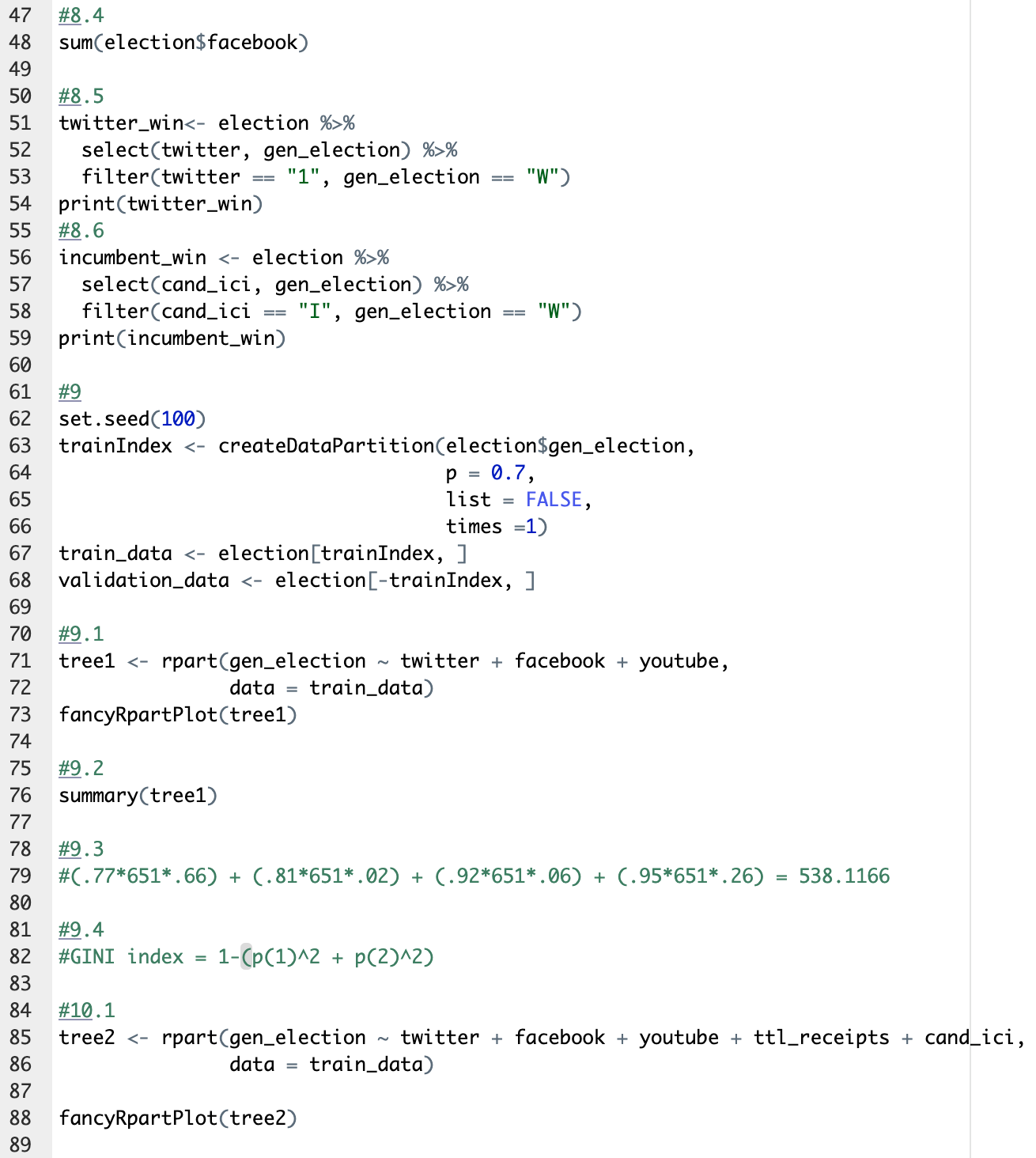
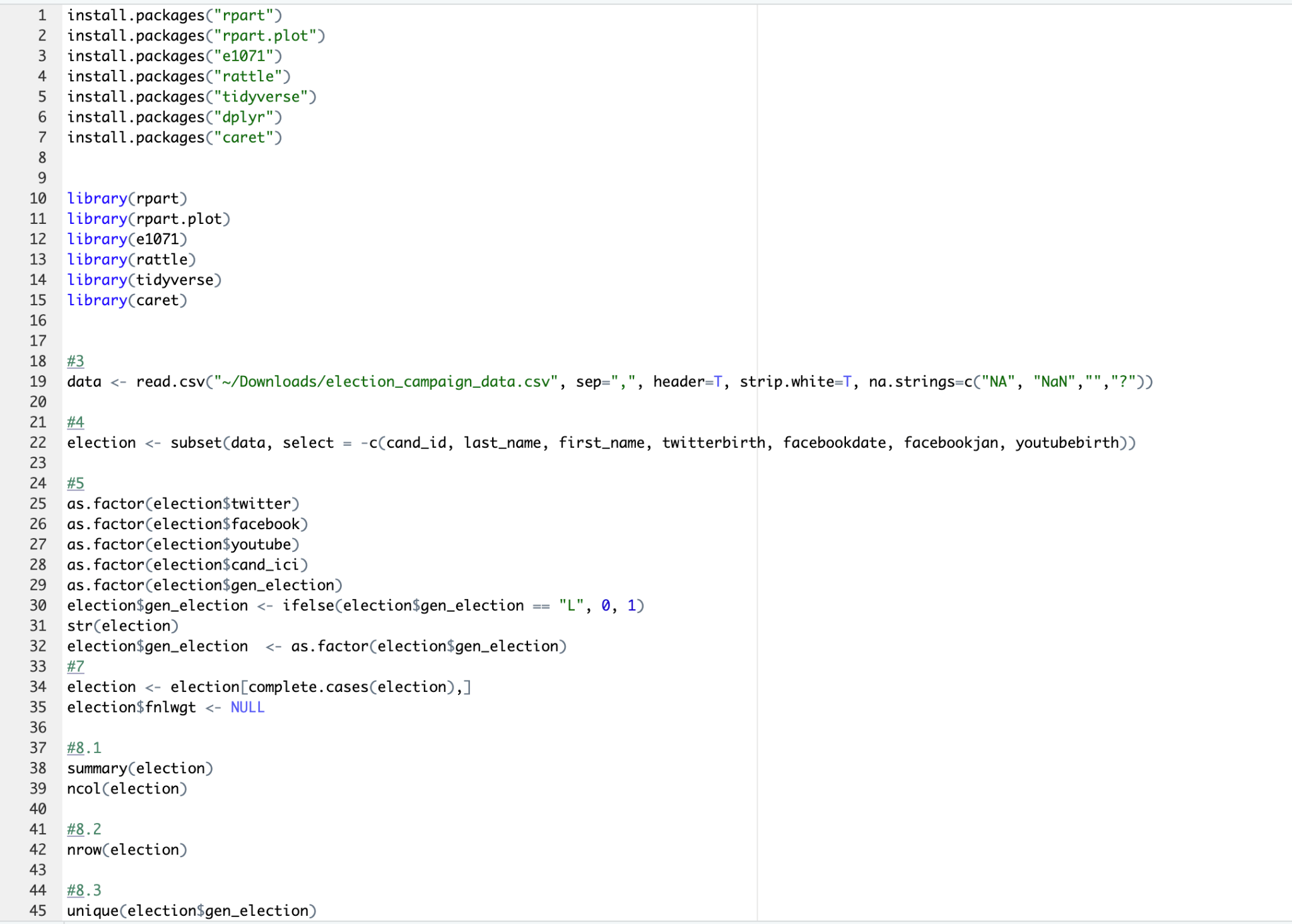
* 1. **(8 points)** Fill out the confusion matrix below using the second decision tree:

|  | Reference | | |
| --- | --- | --- | --- |
| Predicted |  | 0 | 1 |
| 0 | **138** | **21** |
| 1 | **10** | **109** |

1. **(10 points)** Based on your findings, which one is more important? Having social media accounts for the campaign or raising more funds for the campaign? Explain your answer.

It can be determined that raising funds is more important for the campaign as the accuracy for the second decision tree was higher than the first decision tree.

Please don’t forgot to submit your R code (with comments) along with this document.



1. To read more about the general election, please refer to: <http://www.wwnorton.com/college/polisci/campaignsandelections/ch/09/outline.aspx>

   For information about the U.S. Congress, please refer to: <https://en.wikipedia.org/wiki/United_States_Congress> [↑](#footnote-ref-0)