Assignment 3: Advanced Classifiers (100 points)

INFO3237: Business Analytics II

Instructor: Dr. Xue Guo

**Purpose:** To build and test advanced classifiers and prescribe strategies

**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 Assignment3
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 using function mutate\_at(): "twitter","facebook","youtube","cand\_ici","cand\_pty\_affiliation", "gender", "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. More descriptions about the variables can be found in Table 1.
7. Remove all of the observations with any missing values using function drop\_na(). Also use the appropriate code to create a train\_data and validation\_data (70% of original data in the training data and 30% in the testing data).
8. Use package “caret” and train\_data to create a random forest classifier and answer the questions below. Create a random forest using variable gen\_election as the target (class) variable and all of the other variables as predictors (do not include variables such as ID that you think should be excluded from the analysis.).**set.seed = 100**
   1. **(4 points)** Use 5 as the value for ntree and use 2 as the value for mtry. What is the accuracy of the model?

The accuracy is 0.8957445

* 1. **(8 points)** Keep ntree as 5, but change mtry to 7. What is the accuracy of the model? Between 2 and 7, which value of mtry would you recommend?

The accuracy is 0.9200335. Based on the accuracies, I would recommend the 7 value for Mtry as that one has the higher accuracy.

* 1. **(4 points)** Now use the recommended value of mtry but change ntree to 10. What is the accuracy of the model?

The accuracy is 0.9416477

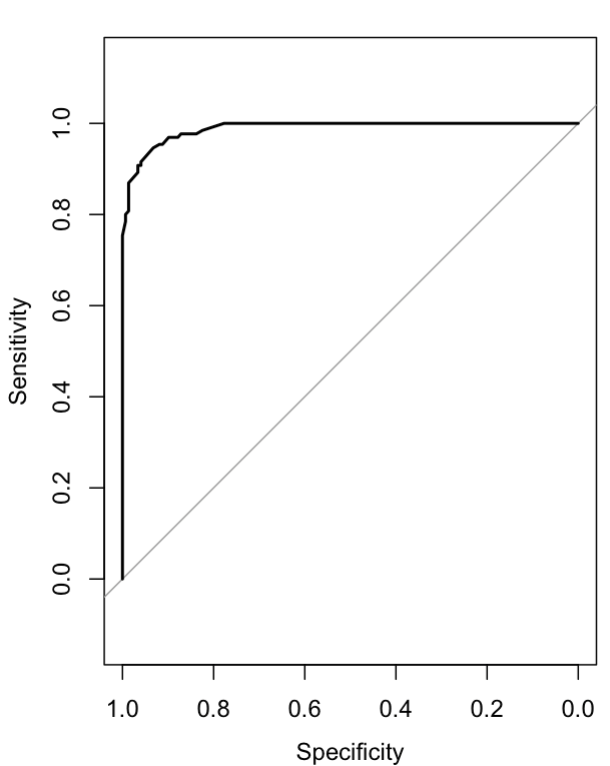
* 1. **(8 points)** Now use the recommended value of mtry but change ntree to 50. What is the accuracy of the model? Between 10 and 50, which value of ntree would you recommend?

The accuracy is 0.9431381 Based on the accuracies, I would recommend the value of 50 for ntree as it has the higher accuracy.

* 1. **(10 points)** Use your recommended ntree and mtry to build the random forest classifier based on train\_data. Now make predictions over validation\_data. Set threshold at 0.5. Create the confusion matrix and paste it in the space below:

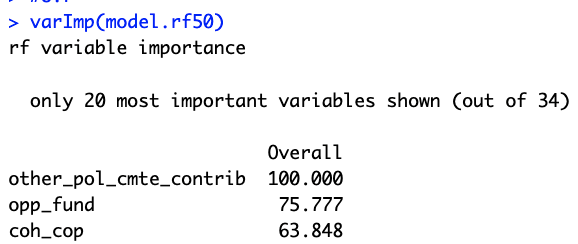
|  | Reference | | |
| --- | --- | --- | --- |
| Predicted |  | 0 | 1 |
| 0 | **145** | **10** |
| 1 | **3** | **120** |

* 1. **(10 points)** Now create the ROC curve and paste in the space below:

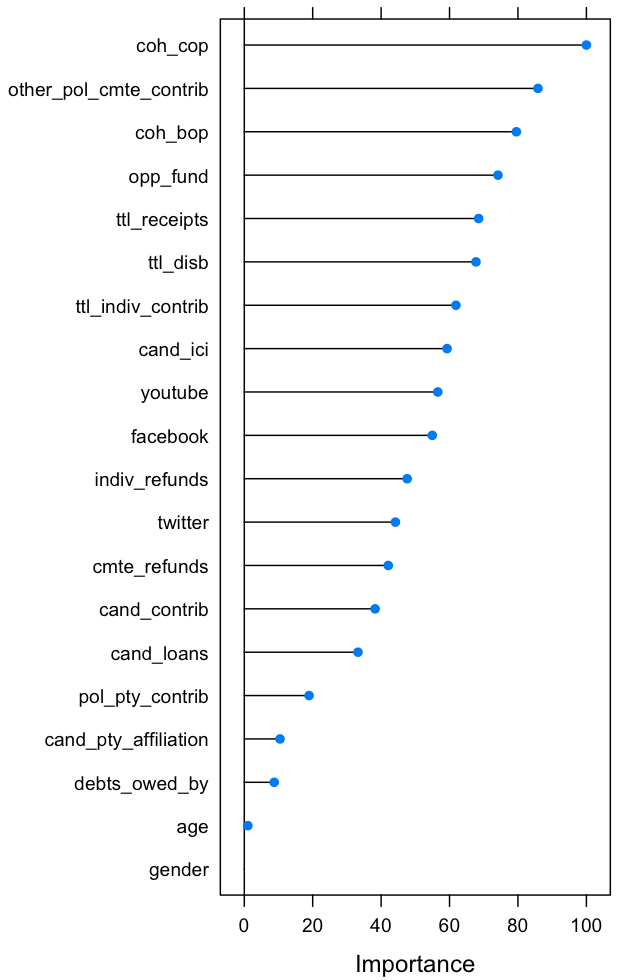
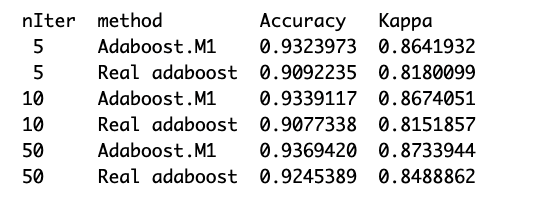


* 1. **(6 points)** Use varImp(rf) to determine the top three variables in predicting gen\_election.

1. other\_pol\_cmte\_contrib 2. opp\_fund 3. Coh\_cop

**

1. Now use boosting method to build a classifier using train\_data. Try the following values of nIter (5,10,50).
   1. **(10 points)** What is the best number of iterations according to the output of boosting model? Show the plot of accuracy vs. iteration and include it below:

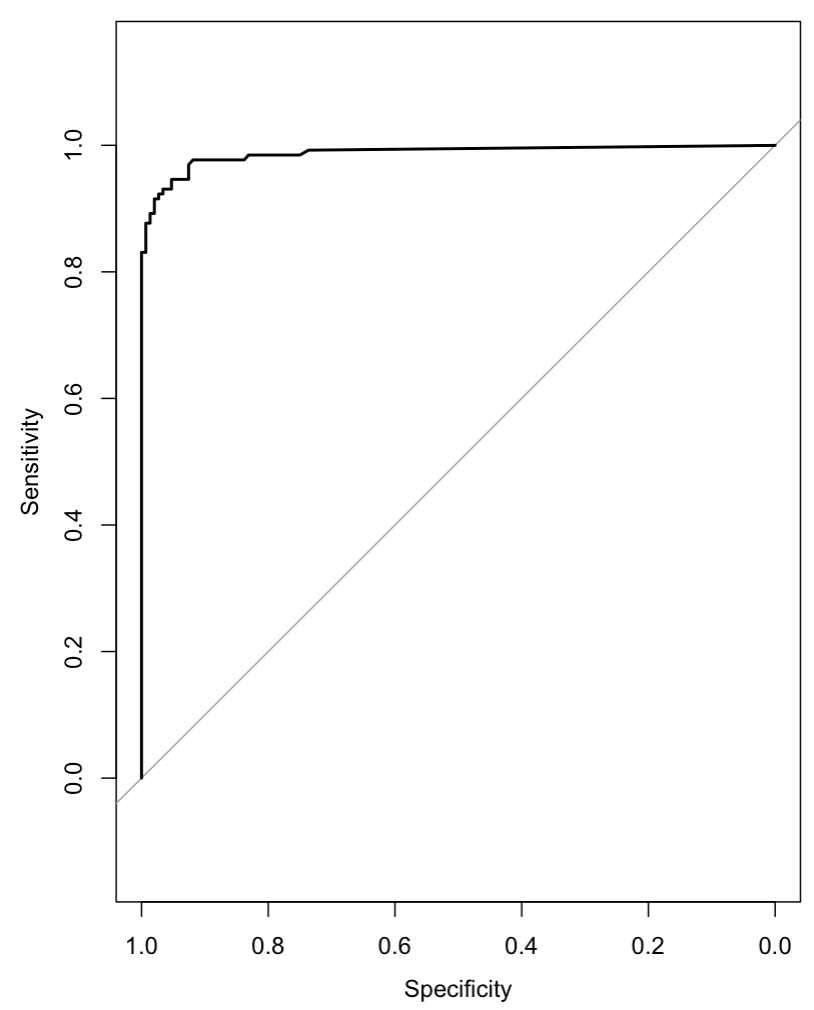


The best number of iterations would be 50 with the adaboost.M1 method.

* 1. **(10 points)** Create the confusion matrix based on validation data:

|  | Reference | | |
| --- | --- | --- | --- |
| Predicted |  | 0 | 1 |
| 0 | **144** | **10** |
| 1 | **4** | **120** |

* 1. **(10 points)** Create the ROC curve and paste in the space below:

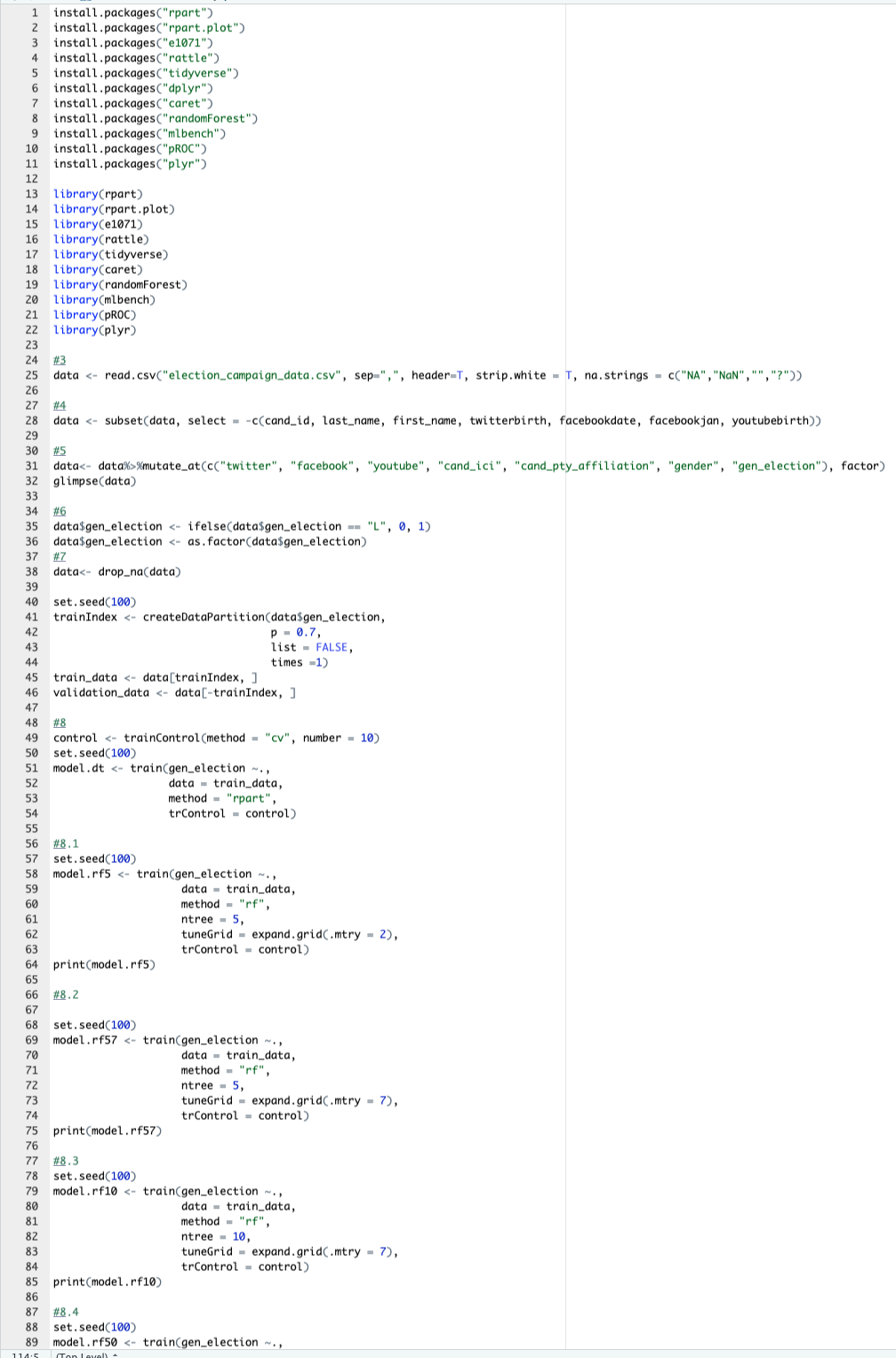


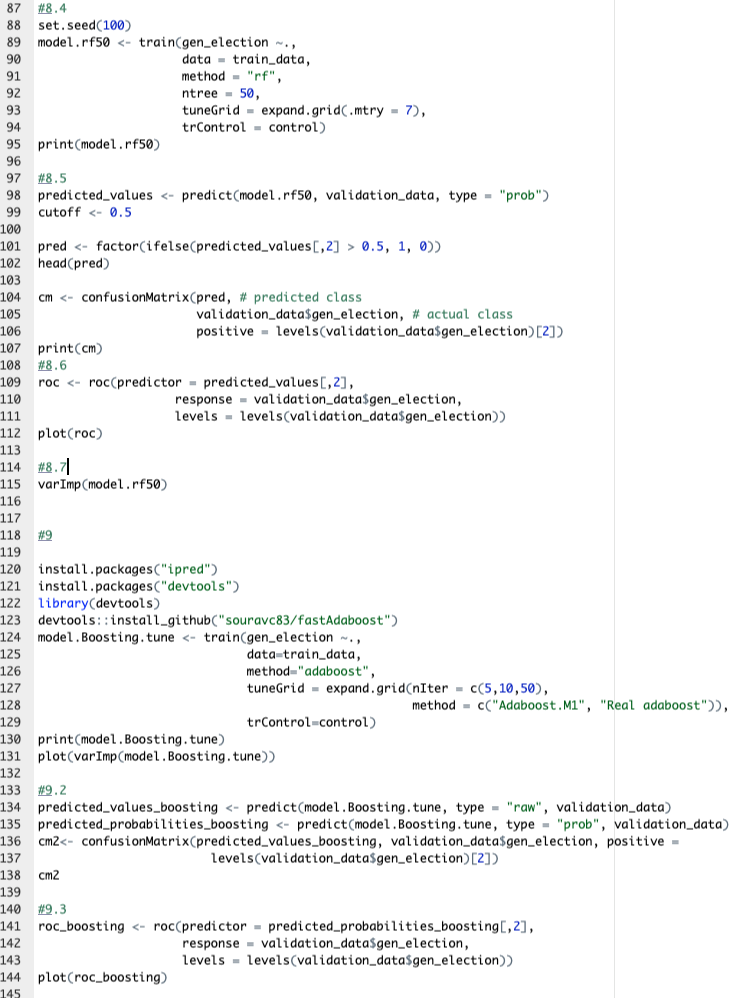
1. (**20 points**) If you were a political consultant, between random forest and boosting method, which one would have you used to make predictions about the elections outcome? Why? If you were giving advice to politicians, you would have encouraged them to take what actions (list at least 2 actions) to help them win the election?

I would advise that the random forest be used as it has a slightly higher accuracy of 94.3% compared to the 93.6% for the boosting method. I would also advise them to focus on the contribution from other political committees as that is the top variable determining the gen\_election variable. This means they should network to gain funds from those committees. This would also in turn help with their ending cash which is also a huge determinant of the target variable.

**Table 1. Key Variables**

| Variable | Description |
| --- | --- |
| cand\_id | A code assigned to a candidate by the FEC |
| last\_name | Last name of candidate |
| first\_name | First name of candidate |
| cand\_ici | Incumbent challenger status with designation: C = Challenger, I = Incumbent, O = Open Seat |
| cand\_pty\_affiliation | The political party affiliation reported by the candidate |
| ttl\_receipts | Total receipts |
| trans\_from\_auth | Transfers from authorized committees |
| ttl\_disb | Total disbursements |
| trans\_to\_auth | Transfers to authorized committees |
| coh\_bop | Beginning cash |
| coh\_cop | Ending cash |
| cand\_contrib | Contributions from candidate |
| cand\_loans | Loans from candidate |
| other\_loans | Other loans |
| cand\_loan\_repay | Candidate loan repayments |
| other\_loan\_repay | Other loan repayments |
| debts\_owed\_by | Debts owed by |
| ttl\_indiv\_contrib | Total individual contribution |
| cand\_office\_st | Candidate's state |
| cand\_office\_district | Candidate's district |
| other\_pol\_cmte\_contrib | Contributions from other political committees |
| pol\_pty\_contrib | Contributions from party committees |
| cvg\_end\_dt | Coverage end date (YYYY-MM-DD) |
| indiv\_refunds | Refunds to individuals |
| cmte\_refunds | Refunds to committees |
| opp\_fund | is the total campaign fund of the opposing candidate |





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)