**GRoup - 02**

**Sachidanandan KC**

**Tarun Kamal Khiani**

**Neelanjan Mukherjee**

**Amareswar N**

**Jyotirmaya Patro**

Machine Learning

Assignment – If during the early part of 2014 (6 months prioir to Elections) you had to predict the upcoming Lok Sabha Elections, what would be your predictive model?

Table of Contents

[Goal 3](#_Toc491763157)

[Solution 3](#_Toc491763158)

[Sources of Data 3](#_Toc491763159)

[Approach 1: What if simulation 5](#_Toc491763160)

[Approach 2: Use candidate Information to predict 2014 results 6](#_Toc491763161)

[Conclusion 8](#_Toc491763162)

# Goal

The goal of this assignment is to the predict the election results 6 months earlier by collating information, churning it and coming up with a model to predict.

# Solution

## Sources of Data

* **Candidate Information**: [www.myneta.info](http://www.myneta.info) provides data about candidates who have participated in General Elections. Candidate Information can be used to form a predictive model.
  + Candidates with serious crimes
    - <http://www.myneta.info/ls2014/index.php?action=summary&subAction=serious_crime>
    - <http://www.myneta.info/ls2009/index.php?action=summary&subAction=serious_crime>
  + Women Candidates
    - <http://www.myneta.info/ls2014/index.php?action=summary&subAction=women_candidate>
    - <http://www.myneta.info/ls2009/index.php?action=summary&subAction=women_candidate>
  + Candidates who won the elections
    - <http://www.myneta.info/ls2014/index.php?action=summary&subAction=winner_analyzed>
    - <http://www.myneta.info/ls2009/index.php?action=summary&subAction=winner_analyzed>
  + Summary Information about the candidates (No. of crimes, Assets)
    - <http://www.myneta.info/ls2014/index.php?action=summary&subAction=candidates_analyzed>
    - <http://www.myneta.info/ls2009/index.php?action=summary&subAction=candidates_analyzed>
* **Constituency & Candidates Information:** We can download the data from <http://eci.nic.in/eci_main1/ElectionStatistics.aspx> > DETAILED RESULTS OF LATEST ELECTIONS (XLS FORMAT) and get information relating to the number for voters, poll percentage, candidates position and their rank in each constituency. However, the data needs to be merged to determine the vote share.

## Approach 1: What if simulation

In this approach, we combine the candidate & elector’s data to determine the vote share for 2009 general elections. Also, we need to take into consideration that Andra Pradesh would be split into Telangana and Andra Pradesh.

Due to poor performance of the UPA lead government and the rise of Narendra Modi we see

* an increase in the vote share of BJP & regional parties – BJP in the north, TRS in Telangana, TDP in Andra Pradesh, TMC/AITC in West Bengal,
* increase in the vote share of AIADMK in Tamil Nadu and reduction of the vote share of DMK and INC in Tamil Nadu
* consistent performance in the rest of south India as earlier
* decrease in the vote share of SP, BSP in Uttar Pradesh
* decrease in the vote share of the UPA parties specially INC

Hence based on a simulation we arrived at the following result:

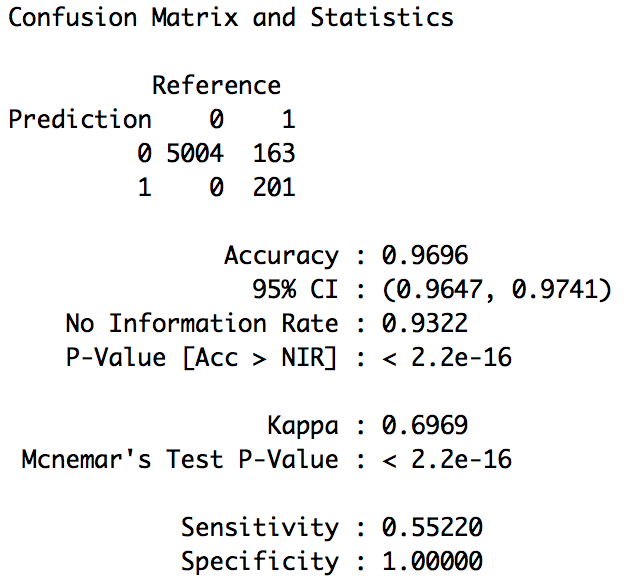
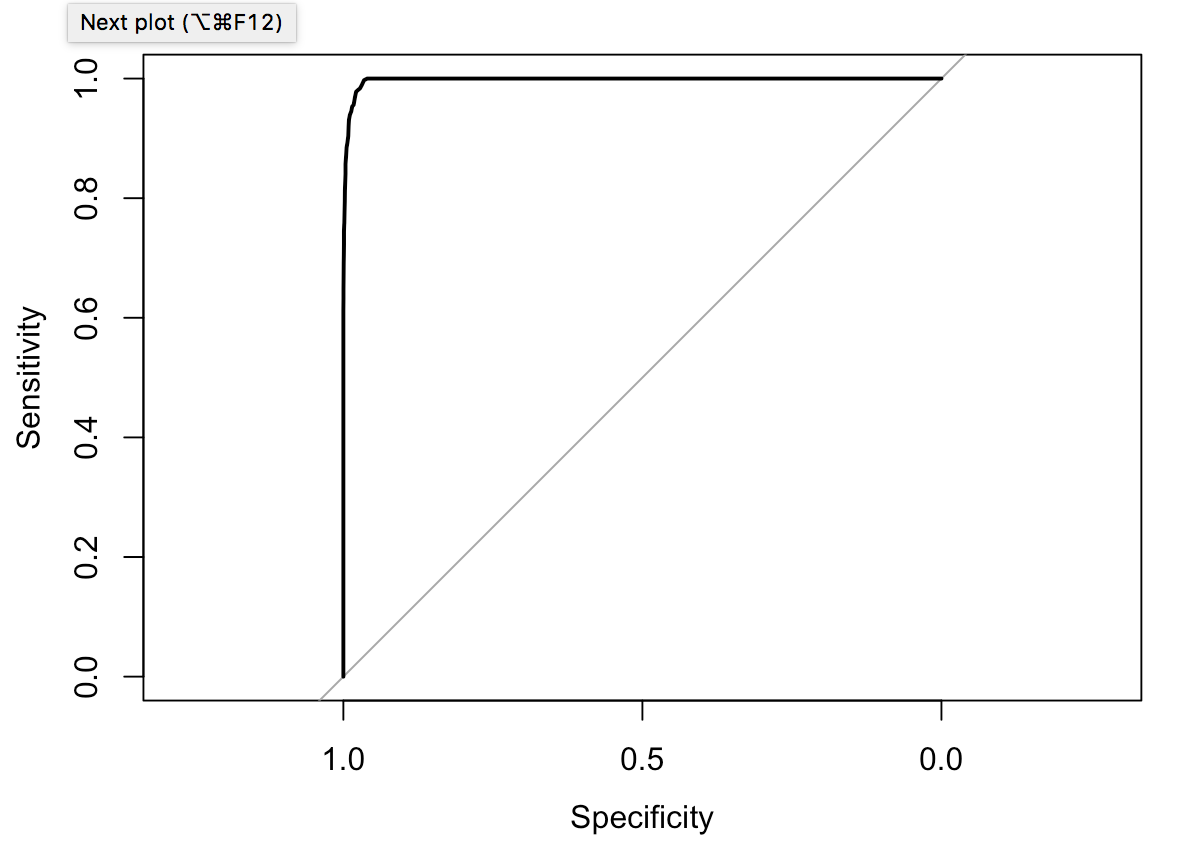
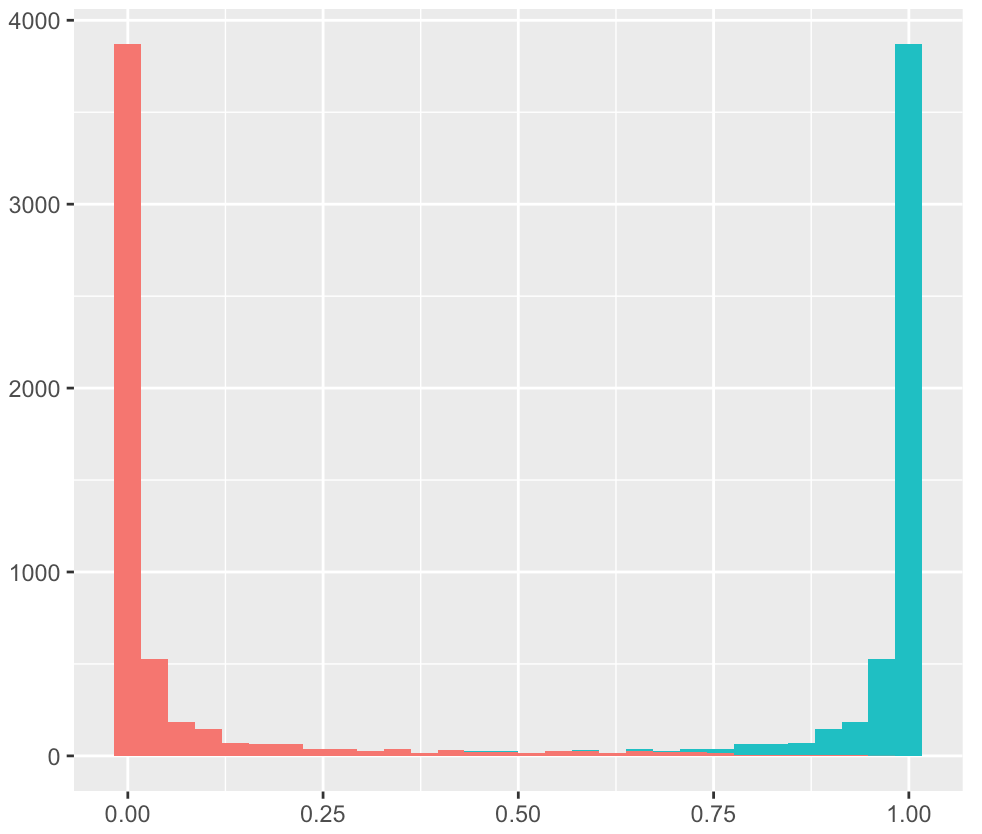
|  |  |
| --- | --- |
| **Party** | **Predicted no. of Seats** |
| BJP | 173 |
| INC | 126 |
| AITC | 24 |
| TDP | 24 |
| JD(U) | 20 |
| SP | 19 |
| ADMK | 17 |
| BJD | 16 |
| DMK | 15 |
| CPM | 14 |

The r script ‘**GE - A what if example.R**’ when run will provide the above result.

## Approach 2: Use candidate Information to predict 2014 results

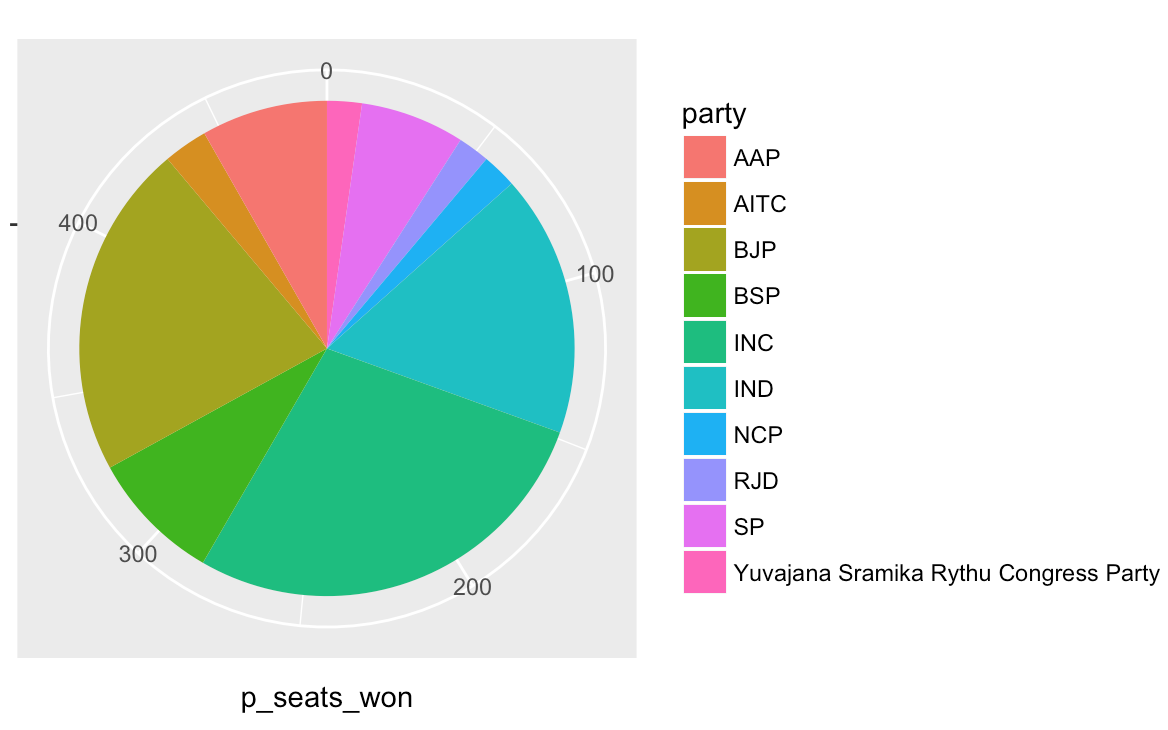
In this approach, we use the candidate information to create a model based on 2009 election results and apply the model to predict 2014 election results. We followed the steps shown below:

1. Used web scrambling to get data about candidates from [www.myneta.info](http://www.myneta.info)
2. Cleansed and merged the following data sets for 2009 & 2014 – Summary Candidate Information (assets, criminal cases), Serious Criminal Cases, Women Candidates, Winning Candidates
3. The r-script ‘**prepareDataForAnalysis.R**’ was used to cleanse, merge & write the cleansed file that would be used to create the model and predict
4. We created a Random Forest to predict if a candidate would win or lose based on the following variables education, assets, seriousCrimes, notSeriousCrimes, gender. The variables like vote share or poll percent would not be available before elections hence omitted.
5. The random forest model worked well on the test data set with an accuracy of 96.96%, AUC of 99.78 %. Hence the model is robust and can be used to predict 2014 election results.



1. We aggregated the predicted 2014 election results based on party and arrived at the following prediction. The prediction as you can see when compared to the actual results is poor as we have not taken other factors into consideration.

|  |  |  |
| --- | --- | --- |
| **Party** | **Seats Won in 2014** | **Predicted Seats Won in 2014** |
| INC | 44 | 135 |
| BJP | 282 | 106 |
| IND | 3 | 83 |
| BSP | 0 | 42 |
| AAP | 4 | 40 |
| SP | 5 | 33 |
| AITC | 34 | 14 |
| NCP | 6 | 11 |
| YSR Congress Party | 9 | 11 |
| RJD | 4 | 10 |

****

1. The r-script ‘predictGE2014.R’ would create the model, validate it with the test data set and predict the election results of 2014.

# Conclusion

* We would suggest going ahead with the What-If-Simulation as candidate information may not be available 6 months prior to the elections. Also, the candidate model does not take into account the mood of the nation and external factors which play a big role.
* Further enhancements are needed and for better interpretation and prediction accuracy. The model can be enhanced by taking into account:
  + PM Candidates across the different alliances
  + Social media analytics
  + Economic Indicators
  + Recent state elections
  + Polling a sample of the population