Project: Statistical Analysis of Assembly Election 2017 for

UTTAR PRADESH

Author: Amit KULKARNI

Project Objective: Data pertaining to candidates who participated the Assembly elections during 2017 has been extracted using the

Election commission of India website: https://eci.gov.in/

And from

myneta.info: http://myneta.info/

This data is filtered for Uttar Pradesh only. Expected to find the significant variables that can impact the outcome of the election and to check if there is any anti-incumbency comparing to 2012 election results.

1. Assumptions: There are no specific assumptions made in the explanation of the solutions
2. Dataset: Election commission of India website: https://eci.gov.in/

   And

   myneta.info: http://myneta.info/

3. **Exploratory Data Analysis:**
   Summary of the merged data is provided explaining the statistical measures

```
                Candidate         Acno                Acname        Ac        Actype     CandidateCategory    CandAge        Party.x
none of the above: 407   Min.   :  1.0   Vishwanathganj:  76   GEN:8520   F   : 593   GEN :6385         40     : 504   IND    :3206
anil kumar       : 270   1st Qu.:106.0   Farrukhabad   :  69   SC :1955   M   :9517   NULL: 407         42     : 421   BSP    : 585
manoj kumar      : 266   Median :198.0   Etah          :  68   ST :  43   NULL: 407   SC  :2816         44     : 407   BJP    : 552
ajay kumar       : 216   Mean   :203.6   Sawaijpur     :  68              O   :   1   ST  :  57         NULL   : 407   RLD    : 500
rakesh kumar     : 210   3rd Qu.:301.0   Paniyara      :  65                          NA's: 853         36     : 402   BMUP   : 467
rajesh kumar     : 168   Max.   :403.0   (Other)       :9319                                            41     : 397   (Other):4355
(Other)          :8981   NA's   :853     NA's          : 853                                            (Other):7980   NA's   : 853
  TotalVotes        Position              Constituency                Party.y      CriminalCase                    Education      TotalAssets
Min.   :    44   Min.   : 1.000   VISHWANATH GANJ:  71   IND               :3875   Min.   : 0.0000   Graduate            :2088   46126.1 :1478
1st Qu.:   522   1st Qu.: 5.000   FARRUKHABAD    :  60   BSP               : 767   1st Qu.: 0.0000   12th Pass           :1984   30,000  :  36
Median :   946   Median : 8.000   SAWAYAZPUR     :  59   BJP               : 710   Median : 0.0000   Post Graduate       :1760   20,000  :  33
Mean   : 12819   Mean   : 8.236   ETAH           :  56   RLD               : 592   Mean   : 0.3644   10th Pass           :1267   1,00,000:  32
3rd Qu.:  2357   3rd Qu.:11.000   BHADOHI        :  55   SP                : 512   3rd Qu.: 0.0000   8th Pass            :1009   2,00,000:  30
Max.   :262741   Max.   :30.000   (Other)        :8739   Bahujan Mukti Party: 481   Max.   :36.0000   Graduate Professional: 853   2,69,000:  28
NA's   :853      NA's   :853      NA's           :1478   (Other)           :3581                     (Other)             :1557   (Other) :8881
   Liabilities   TotalAssets_imp Liabilities_imp CriminalCase_imp Party.y_imp     Education_imp   Actype_imp        Ac_imp         CandAge_imp
 0       :7553   Mode :logical   Mode :logical   Mode :logical    Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 5,00,000:  79   FALSE:10518     FALSE:9040      FALSE:9040       FALSE:9040      FALSE:9040      FALSE:9665      FALSE:9665      FALSE:9665
 6,00,000:  64                   TRUE :1478      TRUE :1478       TRUE :1478      TRUE :1478      TRUE :853       TRUE :853       TRUE :853
 3,00,000:  49
 1,00,000:  45
 2,50,000:  44
(Other)  :2684
```

3.1. **Environmental Set Up and Data Import**: The provided dataset will be analyzed in R STUDIO and MS Excel. Various aspects of the candidate's attributes will be studied and explained

3.2. Install packages and invoke libraries: Variety of packages used and invoked libraries. List given below

3.2.1. rpivotTable
3.2.2. Rpart
3.2.3. Rpartplot

Detailed Working:

Data was gathered in the CSV format from 2 sources viz, election commission of India and the myneta website.

Step 1: Information from election commission of India provided the results from all the states where the Assembly elections were conducted during 2017.

Step 2: Information from Myneta was extracted for Uttar Pradesh only for 2017

Step 3: Data for Uttar Pradesh only was filtered in the election commission of India dataset

Step 4: Both this information was merged in R as shown below

```
setwd ("C:/Users/Amit Kulkarni/Documents/R Programming/Machine Learning/data mining")
data = read.csv("join1.csv")
data1 = read.csv("join2.csv")
fulljoin= merge(data, data1,by = "Candidate", all=T )
View (fulljoin)
```

The data merged basis the candidate name's match

Step 5: To check if there are any NA values in the data. This is done as shown below

```
> colSums(is.na(fulljoin))
        Candidate         StateCode         StateName
                0               853               853
            Month              Year      DistrictName
              853               853               853
             Acno            Acname                Ac
              853               853               853
          Actype CandidateCategory           CandAge
              853               853               853
         Party.x        TotalVotes          Position
              853               853               853
     Constituency           Party.y      CriminalCase
             1478              1478              1478
        Education       TotalAssets       Liabilities
             1478              1478              1478
```

Step 6: NA missing values need to be imputed and it can be done either by calculating the mean, median or the K Nearest Neighbor Imputation function. This function uses the distance calculated from the observation to update the missing values in the dataset. This is shown below

```
library(VIM)
kdata= kNN(Redun_data_rem, variable= c("TotalAssets", "Liabilities","CriminalCase", "Party.y", "Education"), k=3)
```

Party.y	CriminalCase	Education	TotalAssets	Liabilities
BSP	0	12th Pass	3,94,24,827	58,46,335
IND	0	10th Pass	75,106	0
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA

Data before imputing the values

Education	TotalAssets
12th Pass	3,94,24,827
10th Pass	75,106
NA	46126.1

Data after imputation

Step 7: Additionally redundant columns are also removed by subsetting the data

```
reduced_columnsData = subset(data, select = -c(1, 2:7))
```

Step 8: A regression analysis model was run to check the significant variable(s) that would impact the outcome of the election.

There were 6 variables that were studied as shown below. Its adjusted R-squared values are considered to determine the significance level of the variable in relation to the outcome.

Variable Total Assets to Position

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.323 on 6115 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.4464,    Adjusted R-squared:  0.1251
F-statistic: 1.389 on 3549 and 6115 DF,  p-value: < 2.2e-16
```

It can be seen with R-squared value that total assets is not a deciding factor for the position or outcome of the election

Party to the position

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.2 on 9380 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.1985,    Adjusted R-squared:  0.1742
F-statistic: 8.179 on 284 and 9380 DF,  p-value: < 2.2e-16
```

Marginally more that the total assets but does not explain a lot of the position or outcome since the R-squared value is 0.1742

Criminal Case to position

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.596 on 9663 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.01115,   Adjusted R-squared:  0.01104
F-statistic: 108.9 on 1 and 9663 DF,  p-value: < 2.2e-16
```

It can be observed that this variable is insignificant since the r-squared value is just 1%

Education to position

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.584 on 9653 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.01736,   Adjusted R-squared:  0.01624
F-statistic: 15.51 on 11 and 9653 DF,  p-value: < 2.2e-16
```

It can be observed that education is not significant to decide the outcome of the election

Comparison of the caste type to position

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.589 on 9661 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.01458,   Adjusted R-squared:  0.01427
F-statistic: 47.65 on 3 and 9661 DF,  p-value: < 2.2e-16
```

It can be observed that caste is insignificant given that the r-squared value is just 1.4%

Candidate's age to position

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.43 on 9603 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.08687,   Adjusted R-squared:  0.08107
F-statistic: 14.98 on 61 and 9603 DF,  p-value: < 2.2e-16
```

R-squared is higher given the previous 2 variables but still insignificant

Overall 6 identified factors' regression to position

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.836 on 5891 degrees of freedom
  (853 observations deleted due to missingness)
Multiple R-squared:  0.5801,    Adjusted R-squared:  0.3112
F-statistic: 2.157 on 3773 and 5891 DF,  p-value: < 2.2e-16
```

All the 6 factors explain only 31% of the position making the model very weak. 69% still remains unexplained.

Since the provided data could not explain strongly the reason for the outcome of the election there could be some other factors that are impacting.

A quick look at the results of 2012 to 2017 suggests that BJP had a windfall gain in the state. A quick check in MS Excel for 2012 to 2017 results identified that only 60 out of 500+ seats were re-elected. This goes to prove that there was an anti-incumbency for the ruling party in 2012 i.e. Samajwadi Party.

The model suggests us to look for more variables that will be significant to impact the outcome of the election.

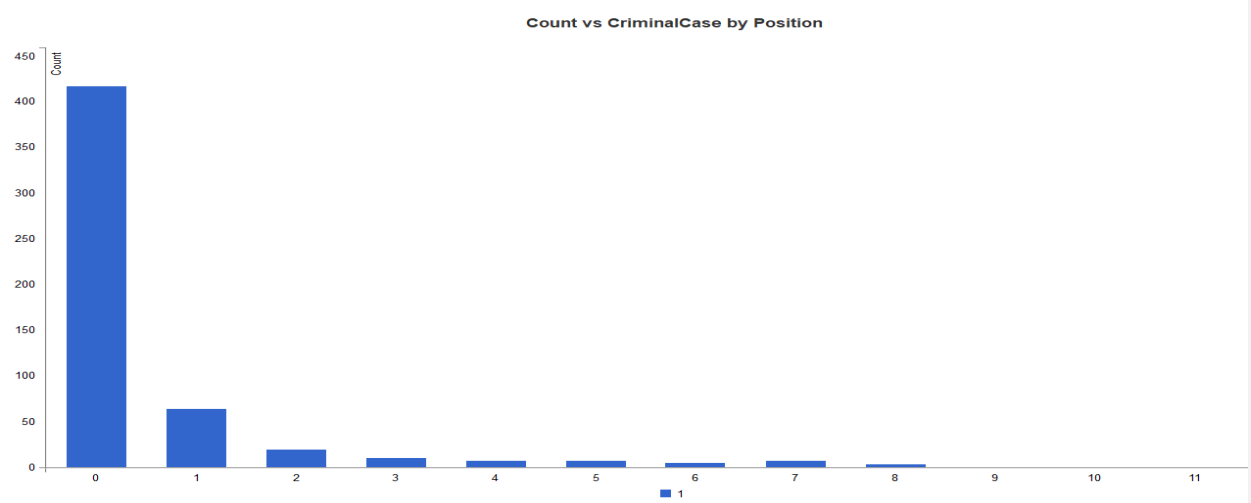
Few other factors that could have impacted would be

1. Modi Wave that went on to topple all other political parties from their ruling state until anti-incumbency was brought by the people in Central India
2. Caste based election could have impacted swaying the things in favor of BJP
3. Expectation of a strong candidate to bring governance and developments within the state etc.

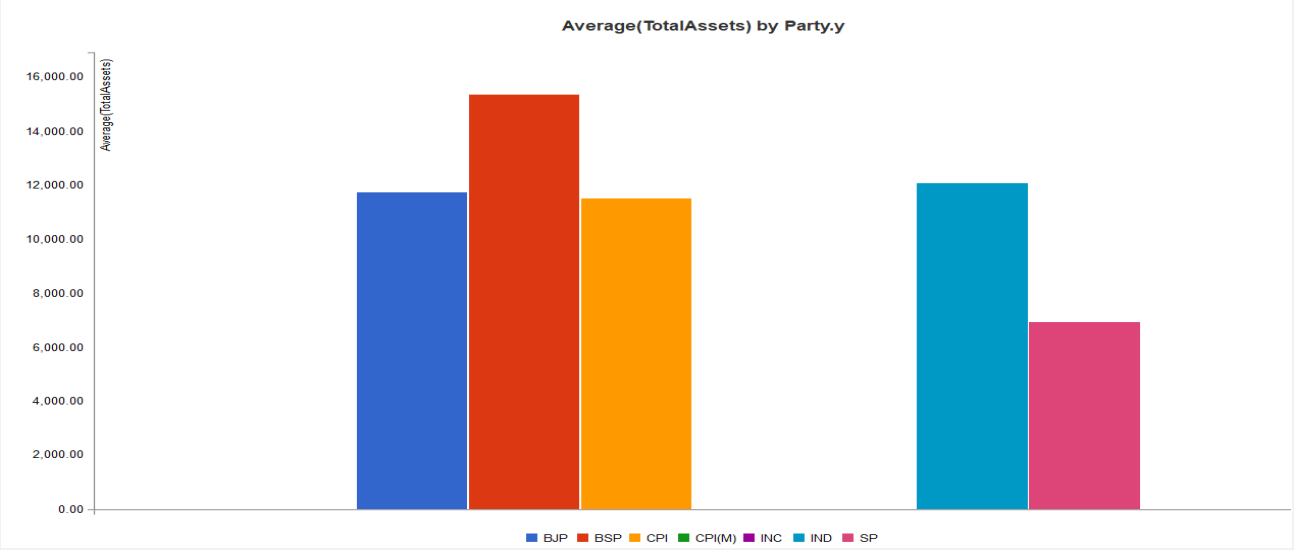
Few other visualization of the available data is provided below

The r-squared values indicated that none of the variables bore impact on the outcome and it is required to do more research to identify the variables that can be considered to relate it to the outcome.

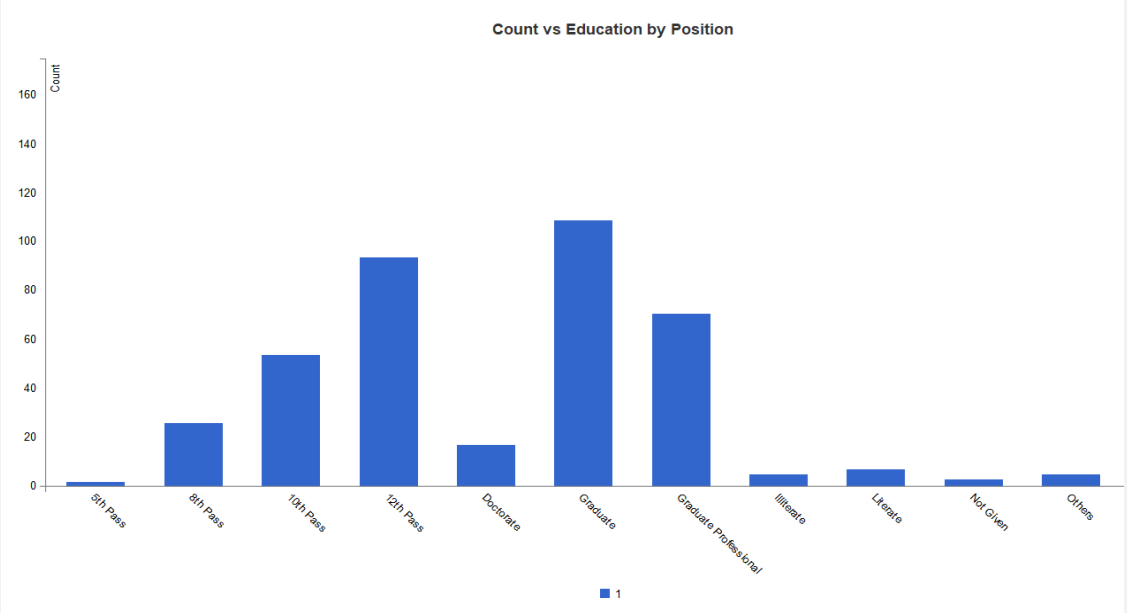
Some graphical representation of the data is done below given multiple parameters



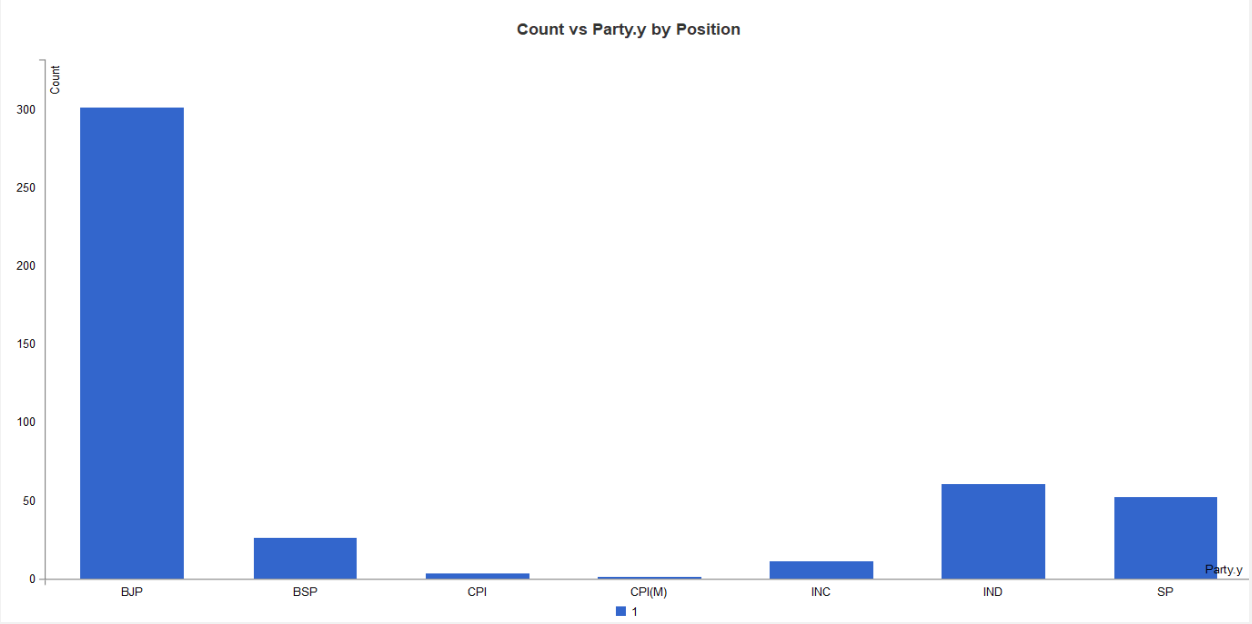
It can be observed from the graph that majority of the candidates who won the election did not have any criminal cases against them. This would mean that people preferred candidates with clean image



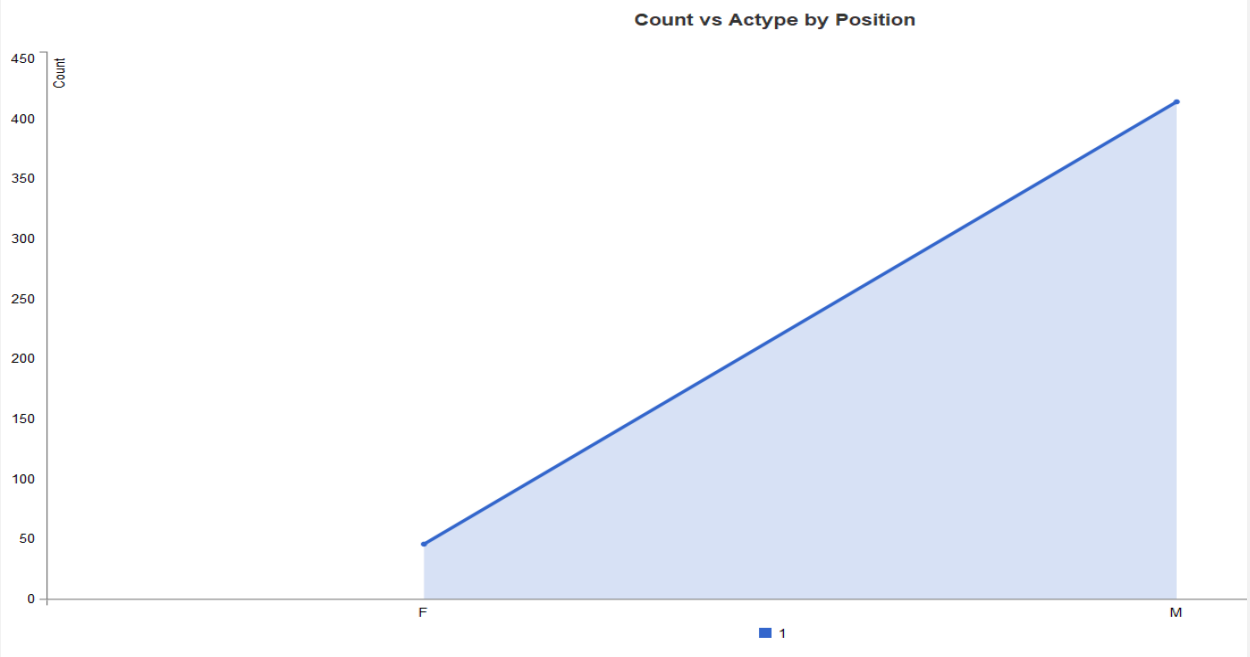
This graph represents the average assets party wise and it can be observed that it is higher for BSP and the rest of the parties are approx. close. The least average assets are for SP



The above graph suggests that people preferred most of the candidates with well-educated backgrounds. There are good number seats where the candidates are graduates and professionals



This graphs represents the overall results party wise for top 7 parties. BJP has a clear mandate in this election. There is a clear mandate for male candidates as shown in the figure.



Graph indicates that the mandate is very clear for male candidates but the proportion of the women participating also needs to be seen before concluding. Percentage success can be checked to see in which direction the mandate swayed.

Concluding remarks: The available dataset was not sufficient to identify significant variables and it demands collection of more variable data.

One thing that is very clear is that there was severe anti-incumbency for SP and which coupled with the Modi wave gave BJP a big boost in the election.