

# Who are the Winners in the 2019 Indian General Election?

## Contents

<b>1. Understanding the Data Set .....</b>	<b>3</b>
<b>2. Exploratory Data Analysis .....</b>	<b>4</b>
2.1 Candidates by Party .....	4
2.2 Number of Constituencies in Each State .....	5
2.3 Winning Rate and Distribution of Education Levels .....	6
2.4 Election Symbols Distribution .....	7
2.5 Age Distribution of Candidates .....	8
2.6 Winning and Losing Candidates by Categories .....	9
2.7 Candidates by Gender .....	10
2.8 Criminal Cases against Candidates .....	10
<b>3. Winner Prediction using Machine Learning Algorithms .....</b>	<b>11</b>
3.1 Decision Tree .....	11
3.2 Random Forest .....	13
3.3 Logistic Regression .....	14
<b>4. Conclusion .....</b>	<b>15</b>

The Lok Sabha, the lower house of the Parliament of India, is made up of Members of Parliament (MPs). Each MP, represents a single geographic constituency. There are currently 543 constituencies. The 17th Lok Sabha Election held in 2019 from April to May 2019 in 7 phase. Around 910 Million voters were eligible to vote and the voter turnout was 67%, this was the highest ever voting recorded by Election Commission of India. Voting percent of women's voters were also increased than previous Lok Sabha elections.

In this project, we will examine what attributes make a candidate successful in this election, and try to model the mechanism using supervised machine learning algorithms.

## 1. Understanding the Data Set

The dataset could be downloaded from Kaggle dataset with the link <https://www.kaggle.com/prakrutchauhan/indian-candidates-for-general-election-2019>. There are 19 variables in the original dataset: STATE, CONSTITUENCY, NAME, WINNER, PARTY, SYMBOL, GENDER, CRIMINAL CASES, AGE, CATEGORY, EDUCATION, ASSETS, LIABILITIES, GENERAL VOTES, POSTAL VOTES, TOTALVOTES, OVER TOTAL ELECTORS IN CONSTITUENCY, OVER TOTAL VOTES POLLED IN CONSTITUENCY, TOTAL ELECTORS. Using “summary” to display the properties of the dataset:

```
> summary(df)

      STATE          CONSTITUENCY        NAME      WINNER      PARTY      SYMBOL
Uttar Pradesh : 274  AURANGABAD : 14  NOTA       : 245  Min.    :0.0000  BJP   :420  Lotus   :420
Bihar         : 244  GAYA (SC)  : 12  Ajay Kumar  : 2  1st Qu.:0.0000  INC   :413  Hand    :413
Tamil Nadu   : 217  MAHARAJGANJ:  9  ATUL KUMAR SINGH: 2  Median  :0.0000  NOTA  :245
West Bengal   : 193  UJIARPUR   : 9  Rahul Gandhi  : 2  Mean    :0.2382  IND   :201  Elephant :166
Maharashtra  : 192  ARUKU     : 8  SANJAY KUMAR  : 2  3rd Qu.:0.0000  BSP   :163  Bicycle  : 65
Andhra Pradesh: 121  BARAMULLA :  8  SURENDRA RAM  : 2  Max.   :1.0000  CPI(M) :100  Hammer, sickle and star: 63
(Other)       :1022 (Other)    :2023 (Other)    :2008 (Other)  :721 (Other)  :891
                                         (Other)    :2008 (Other)  :721 (Other)  :891

      GENDER      CRIMINAL.CASES      AGE      CATEGORY      EDUCATION      ASSETS
      : 245      0 :1242  Min.  :25.00      : 245 Post Graduate  :502      : 245
FEMALE: 258      1 :313  1st Qu.:43.25  GENERAL:1392 Graduate  :441 Not Available  : 22
MALE :1760      2 :245  Median :52.00  SC      : 383 Graduate Professional:336 Nil      : 3
      2 :119  Mean   :52.27  ST      : 243 12th Pass    :256 Rs 1,75,000\n ~ 1 Lacs+  : 2
      3 :104  3rd Qu.:61.00          :       :       :245 Rs 1,93,54,59,756\n ~ 193 crore+:  2
      4 : 64  Max.   :86.00          :       :       :196 Rs 11,95,43,561\n ~ 11 crore+:  2
(Other): 176 NA's  :245          :       :       :287 (Other)  :1987
                                         (Other)  :287 (Other)  :1987

      LIABILITIES      GENERAL.VOTES      POSTAL.VOTES      TOTAL.VOTES      OVER.TOTAL.ELECTORS..IN.CONSTITUENCY
      : 634  Min.  : 1339 Min.  : 0.0 Min.  : 1342 Min.  : 0.09794
      : 245  1st Qu.: 21035 1st Qu.: 57.0 1st Qu.: 21163 1st Qu.: 1.29652
Not Available  : 22 Median : 153934 Median : 316.0 Median : 154489 Median :10.51055
Rs 5,00,000\n ~ 5 Lacs+: 10 Mean   : 261599 Mean   : 990.7 Mean   : 262590 Mean   :15.81141
Rs 1,00,000\n ~ 1 Lacs+:  8 3rd Qu.: 485804 3rd Qu.: 1385.0 3rd Qu.: 487232 3rd Qu.:29.46818
Rs 50,000\n ~ 50 Thou+:  8 Max.   :1066824 Max.   :19367.0 Max.   :1068569 Max.   :51.95101
(Other)       :1336

OVER.TOTAL.VOTES.POLLED..IN.CONSTITUENCY TOTAL.ELECTORS
Min.  : 1.000 Min.  : 55189
1st Qu.: 1.899 1st Qu.:1530014
Median :16.222 Median :1679030
Mean   :23.191 Mean   :1658016
3rd Qu.:42.590 3rd Qu.:1816857
Max.   :74.412 Max.   :3150313
```

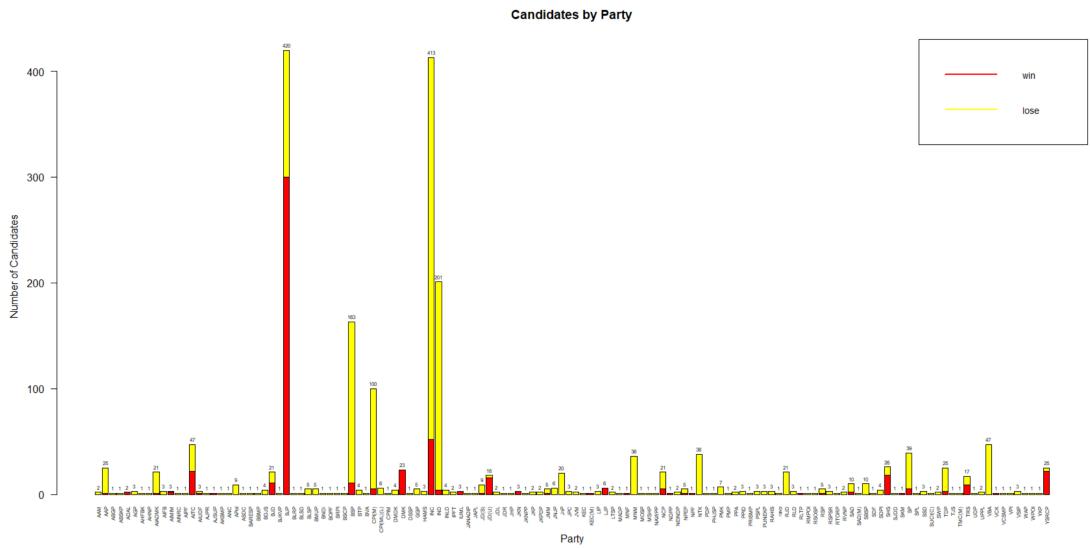
The Minimum age of the candidates was 25 whereas maximum age was 86. Average age of all the candidates who contested election was 52. As can be seen from the CATEGORY column, most candidates are GENERAL. General Category, also referred as Forward caste/General Class, is a term used in India to denote social groups that is ahead of other Indians on an average on economic and social front. Forward castes form about 18.8% of the population, the number varying by region.

"None of the Above" (or NOTA) has been provided as an option to the voters of India in most elections since 2019. By expressing a preference for none of the above, a citizen can choose not to vote for any candidates who are contesting the elections. The vote does not hold any electoral value: even if a majority of votes were cast for NOTA, the candidate with the largest vote share would still be the winner. NOTA enables the voter to show their unacceptance for the fielded candidates. There are 245 NOTA (None of The Above) in NAME, and after we exclude them, there were 2018 candidates who contested 2019 LokSabha Election. Besides, after excluding NOTA, there is no missing values in the dataset anymore.

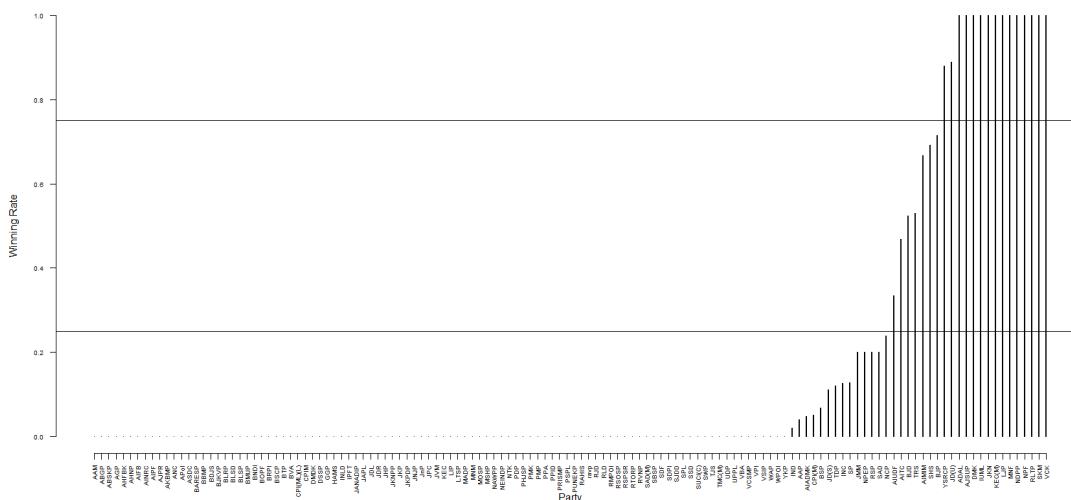
STATE		CONSTITUENCY		NAME		WINNER		PARTY		SYMBOL	
Uttar Pradesh : 251	AURANGABAD :	13	Ajay Kumar	:	2	Min. : 0.0000	BJP	: 420	Lotus	:	420
Bihar : 218	GAYA (SC) :	11	ATUL KUMAR SINGH	:	2	1st Qu.: 0.0000	INC	: 413	Hand	:	413
Tamil Nadu : 189	CHATRA :	8	Rahul Gandhi	:	2	Median : 0.0000	IND	: 201	Elephant	:	166
Maharashtra : 175	MAHARAJGANJ :	8	SANJAY KUMAR	:	2	Mean : 0.2671	BSP	: 163	Bicycle	:	65
West Bengal : 173	SHEOHAR :	8	SURENDRA RAM	:	2	3rd Qu.: 1.0000	CPI(M)	: 100	Hammer, Sickle and Star:	63	
Andhra Pradesh: 101	SUPAUL :	8	A N RADHAKRISHNAN	:	1	Max. : 1.0000	AITC	: 47	Cup & Saucer	:	52
(Other): 911	(Other)	1962	(Other)	:	2007			(Other): 674	(Other)		: 839
GENDER		CRIMINAL.CASES		AGE		CATEGORY		EDUCATION		ASSETS	
:	0	0	: 1242	Min. : 125.00		:	0	Post Graduate	: 502	Not Available	
FEMALE: 258	1	: 313	1st Qu.: 43.25	GENERAL:	1392	Graduate	: 441	NIL		:	3
MALE : 1760	2	: 119	Median : 52.00	SC :	383	Graduate Professional:	336	Rs 1,75,000\n ~ 1 Lacs+		:	2
	3	: 104	Mean : 52.27	ST :	243	12th Pass	: 256	Rs 1,93,54,59,756\n ~ 193 Crore+		:	2
	4	: 64	3rd Qu.: 61.00			10th Pass	: 196	Rs 11,95,43,561\n ~ 11 Crore+		:	2
	5	: 42	Max. : 86.00			8th Pass	: 78	Rs 15,88,77,063\n ~ 15 Crore+		:	2
	(Other): 134					(Other)	: 209	(Other)			: 1985
LIABILITIES GENERAL.VOTES POSTAL.VOTES TOTAL.VOTES OVER.TOTAL.ELECTORS..IN.CONSTITUENCY											
Rs 0\n ~		: 634	Min. : 1339	Min. :	0	Min. : 1342	Min. :	0.09794			
Not Available		: 22	1st Qu. : 30476	1st Qu. :	97	1st Qu. : 30744	1st Qu. :	1.95362			
Rs 5,00,000\n ~ 5 Lacs+:	10	Median : 284630	Median :	463	Median :	285525	Median :	18.03686			
Rs 1,00,000\n ~ 1 Lacs+:	8	Mean : 291190	Mean :	1105	Mean :	292295	Mean :	17.59681			
Rs 50,000\n ~ 50 Thou+ :	8	3rd Qu. : 505862	3rd Qu. :	1546	3rd Qu. : 507618	3rd Qu. :	30.70811				
Rs 3,00,000\n ~ 3 Lacs+:	7	Max. : 1066824	Max. :	19367	Max. :	1068569	Max. :	51.95101			
(Other)		: 1329									
OVER.TOTAL.VOTES.POLLED..IN.CONSTITUENCY TOTAL.ELECTORS											
Min. : 1.000			Min. :	55189							
1st Qu. : 2.871			1st Qu. :	1530404							
Median : 27.750			Median :	1679891							
Mean : 25.808			Mean :	1660261							
3rd Qu. : 44.350			3rd Qu. :	1823404							
Max. : 74.412			Max. :	3150313							

## 2. Exploratory Data Analysis

## 2.1 Candidates by Party

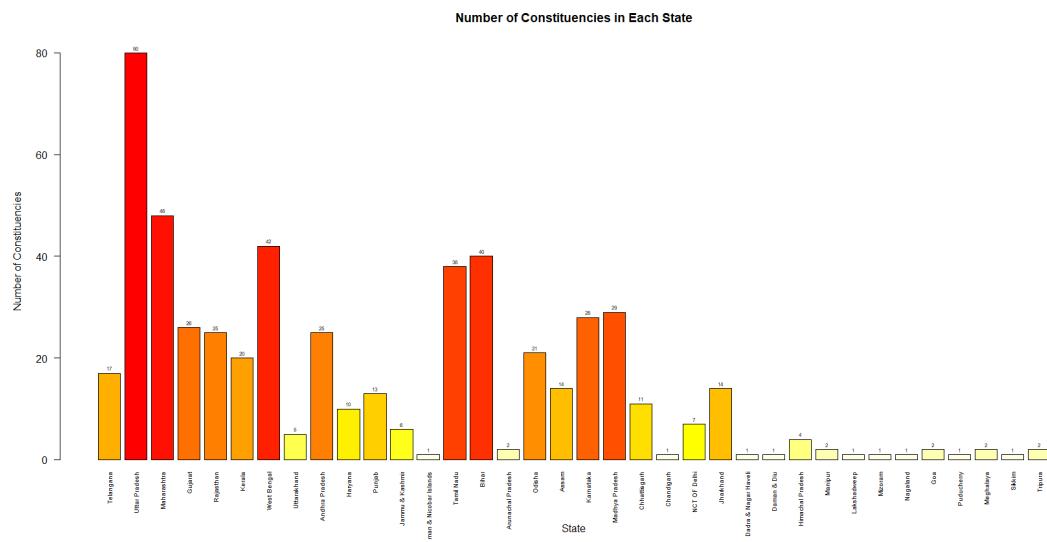


India has a multi-party system with recognition accorded to national and state and district level parties. There were 133 parties involving in 2019 Lok Sabha election. BJP had the largest number, 420, of candidates contesting the election, followed by INC 413 candidates. However, BJP achieved a landslide victory winning 300 seats out of 539 seats whereas INC won only 52 seats. Candidates in the party INC had much lower winning rate than in the BJP. Some parties, such as DMK, YSRCP, LJP, etc., had very high winning rates in spite of low numbers of candidates.

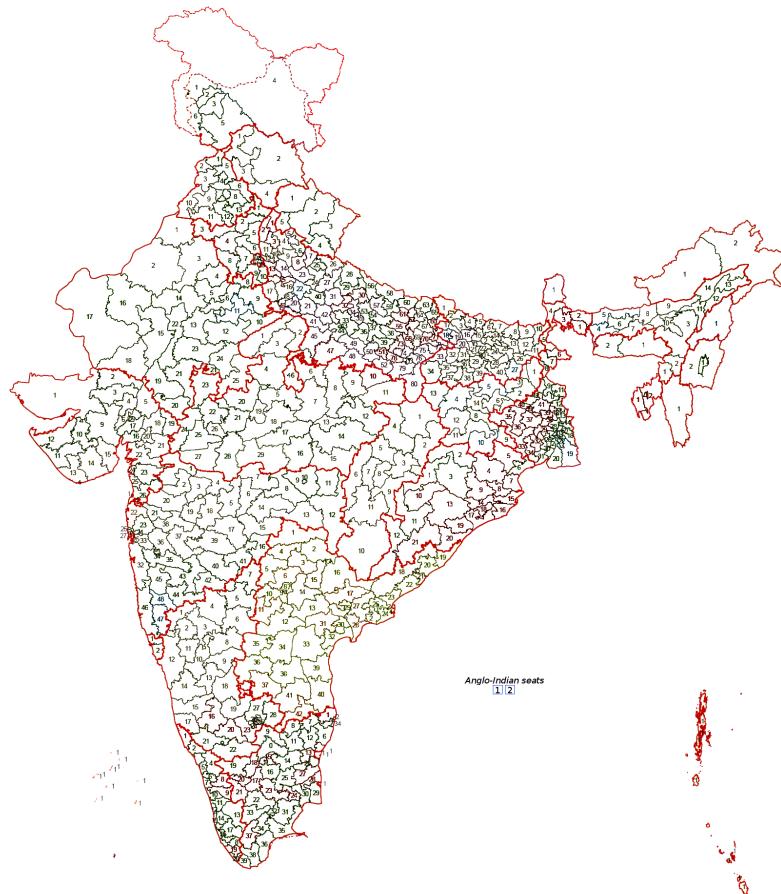


In order to consider parties as a factor in predicting winning rates, we classify 133 parties into three categories based on their winning rates. Winning rates higher than 75% (The third quantile), higher than 25% and lower than and equal to 75%, and lower than 25% (The first quantile) are classified as “high”, “medium”, and “low” respectively. As a result, there are 88 parties in “high”, 1393 in “medium”, and 537 in “low”.

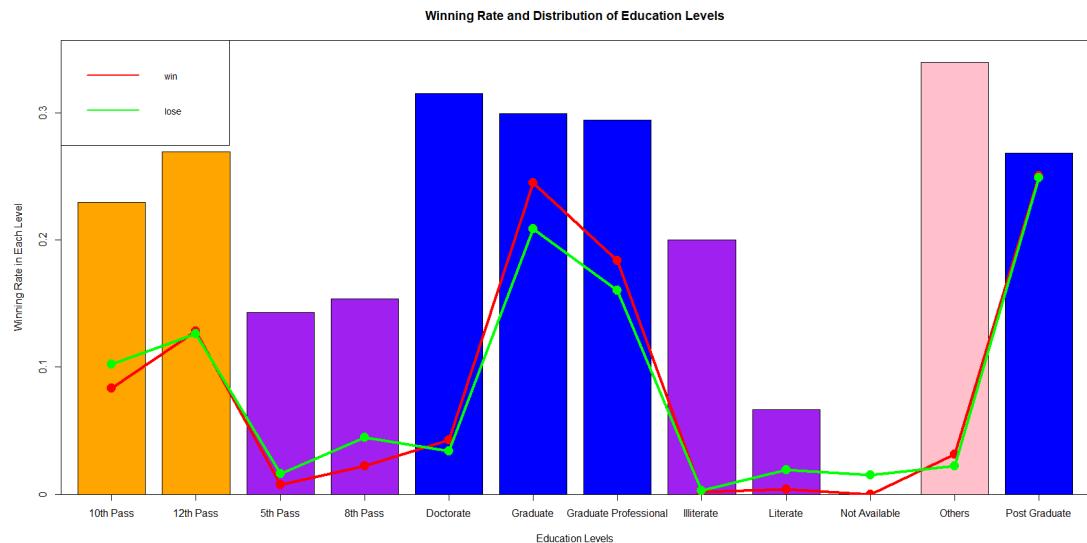
## 2.2 Number of Constituencies in Each State



There are 36 states and 539 constituencies shown in the data set. India is a federal union comprising 28 states and 8 union territories, for a total of 36 entities. There are 5 categories of states and union territories: 5 Very Big, 7 Big, 6 Average, 4 Small, rest 14 are categorized as Tiny. The constituencies are delimited in a manner that the population of each Parliamentary and Assembly Constituency in a State so far as practicable be the same throughout the State. The following graph shows the delineation of constituencies across India.



### 2.3 Winning Rate and Distribution of Education Levels



Based on the education level, we divide 12 levels of education into 5 groups and use different colors to distinguish them. There are 1353 candidates possessing advanced education (Adv\_edu) including Doctorate, Graduate, Graduate Professional, Post Graduate; 452 candidates having high level education (High\_edu) including 10<sup>th</sup> Pass and 12<sup>th</sup> Pass; 163 candidates with low level education (Low\_edu) including 5<sup>th</sup> Pass, 8<sup>th</sup> Pass, Literate, Illiterate, Not Available; and Others (Others\_edu). It can be seen that candidates running for India's lower house of parliament are generally highly educated.

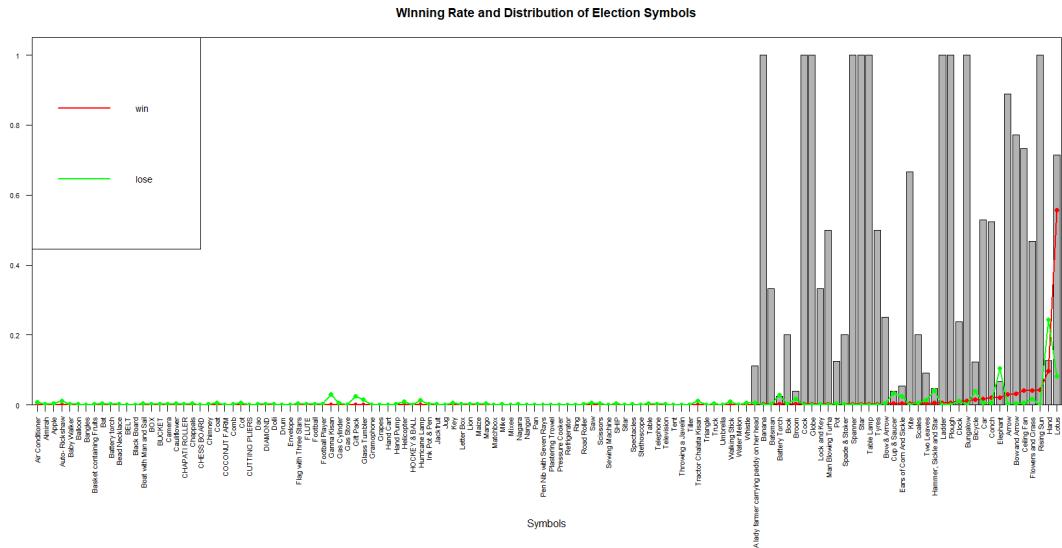
Showed in the Figure, people with Others or graduate level of education had relatively higher winning rates whereas those with a low level of education had a low probability of winning. Though people with graduate levels of education composed the highest proportion of winning and losing candidates, a higher percentage of the winning candidates were graduate, graduate professional and doctorate, and a higher percentage of the losing candidates had lower educational levels such as 10<sup>th</sup> Pass, 5<sup>th</sup> pass, 8<sup>th</sup> pass, Literate, and Not Available.

#### 2.4 Election Symbols Distribution

When India held its first national election in 1951-52, almost 3 in 4 voters were illiterate. To help them identify the party of their choice, visual symbols were allotted to parties and candidates. While literacy levels have increased dramatically since, the abiding appeal of party symbols has not faded.

All registered parties contesting elections need to choose a symbol from a list of available symbols offered by the Election Commission of India. If a party is recognized as a national or state party, its symbol is reserved for its exclusive use in the country or in the state. For example, India's national flower, the lotus, is the symbol of the ruling Bharatiya Janata Party, or BJP. The lotus is associated with the Hindu goddess of knowledge and symbolizes the party's link to Hinduism and its traditions. While India's seven national parties and 64 state parties have fixed symbols, the Election Commission also has a pool of "free" symbols that can be used by thousands of smaller, lesser-known organizations.

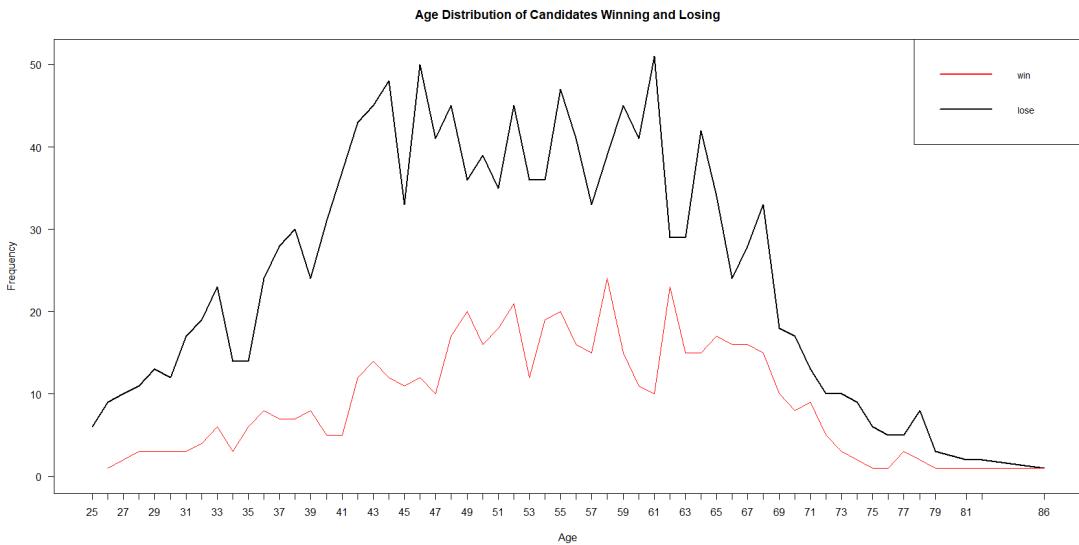




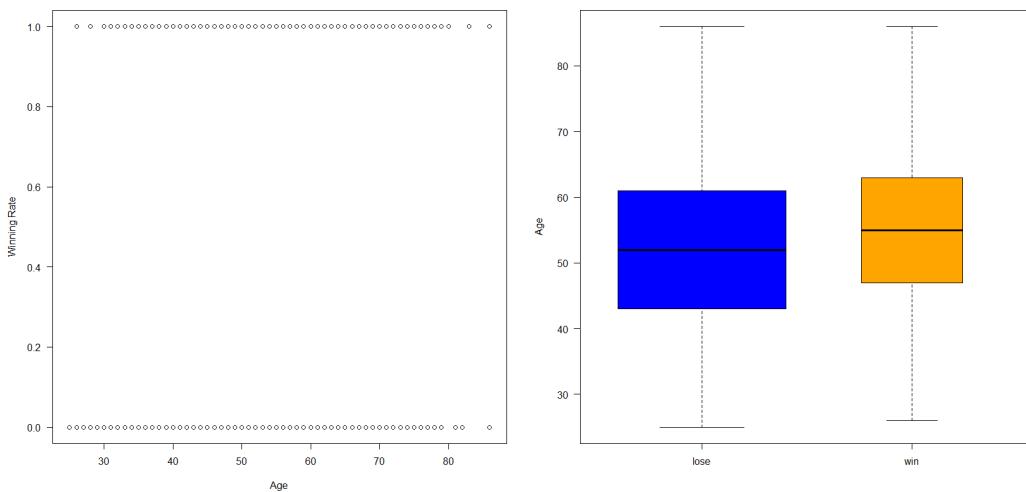
There were 127 symbols contested in the 2019 Lok Sabha, but only a few parties won seats. Among the winning candidates, the biggest share came from the lotus party, the BJP, followed by the Indian National Congress, symbolized by the hand. However, Indian National Congress also had the highest number of losing candidates as reflected by the highest green point in the figure. The fact also shows in the low bar chart of hand symbol party.

Disregarding those symbols that had 100 percent winning rate because only one or two candidates ran and won, candidates in the parties symbolized by lotus, rising sun, flowers and grass, ceiling fan, bow and arrow, and arrow had a higher probability of winning.

### **2.5 Age Distribution of Candidates**



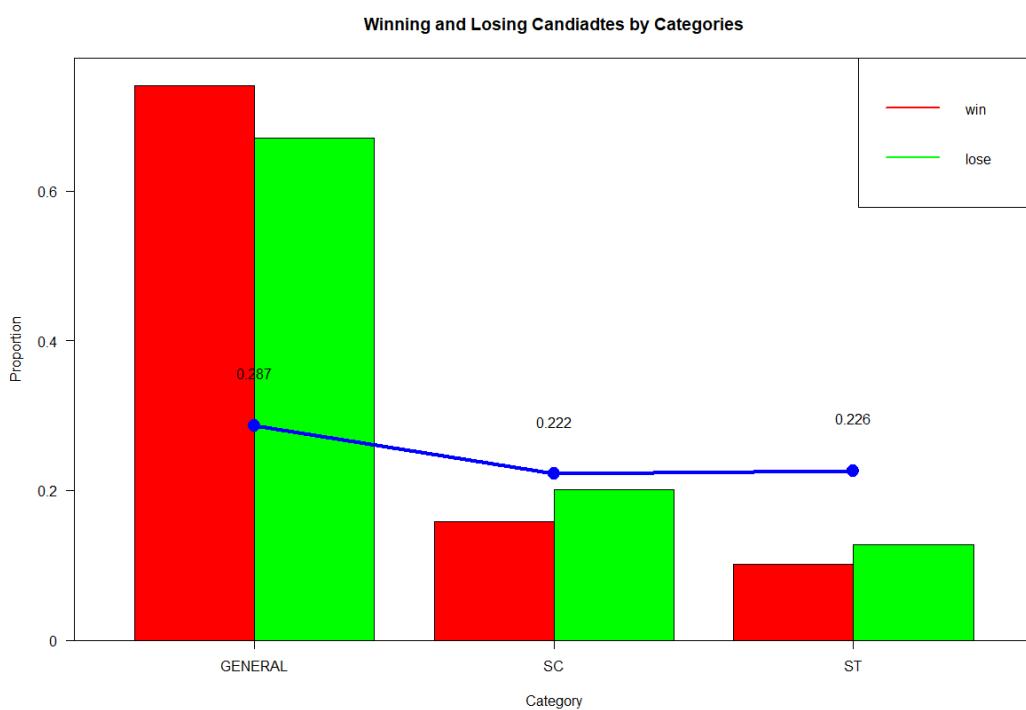
The winning and losing candidates had very similar age distribution, with most being between 40 and 60 years old.



According to the scatter plot and the boxplot, age cannot differentiate winning rate. Therefore, we do not consider age as a predictor in our prediction models.

## 2.6 Winning and Losing Candidates by Categories

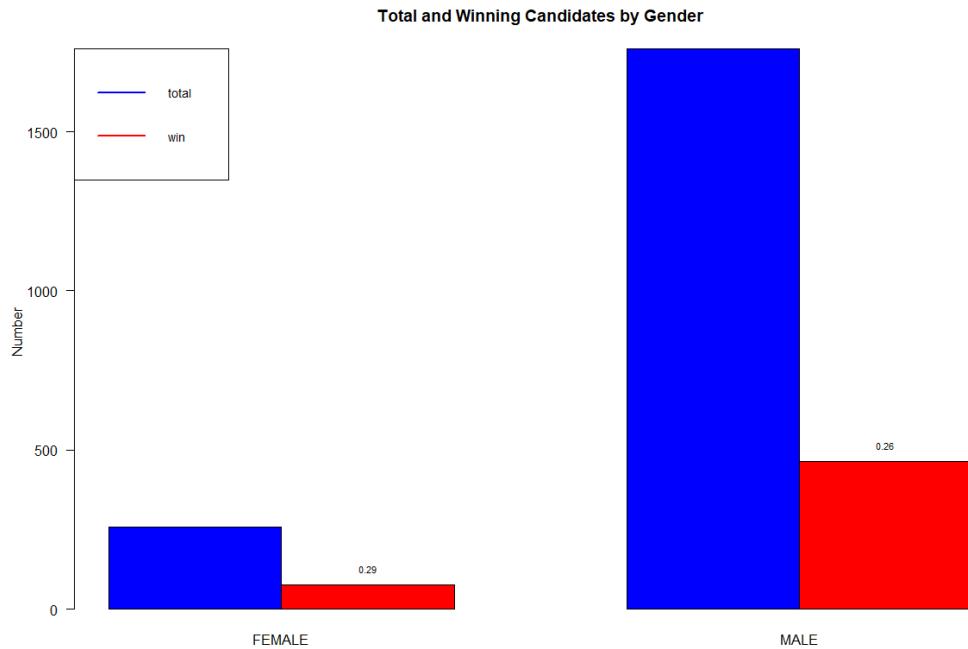
The Scheduled Caste (SCs) and Scheduled Tribes (STs) are officially designated groups of people in India. For much of the period of British rule in the Indian subcontinent, they were known as the Depressed Classes. In modern literature, the Scheduled Castes are sometimes referred to as Dalit, meaning "broken/scattered" in Sanskrit. The Scheduled Castes and Scheduled Tribes comprise about 16.6% and 8.6%, respectively, of India's population (according to the 2011 census). The Constitution (Scheduled Castes) Order, 1950 lists 1,108 castes across 29 states in its First Schedule, and the Constitution (Scheduled Tribes) Order, 1950 lists 744 tribes across 22 states in its First Schedule. The Constitution lays down the general principles of positive discrimination for SCs and STs. Since the independence of India, the Scheduled Castes and Scheduled Tribes were given Reservation status, guaranteeing political representation.



The highest proportion of the winning and losing candidates was of the General

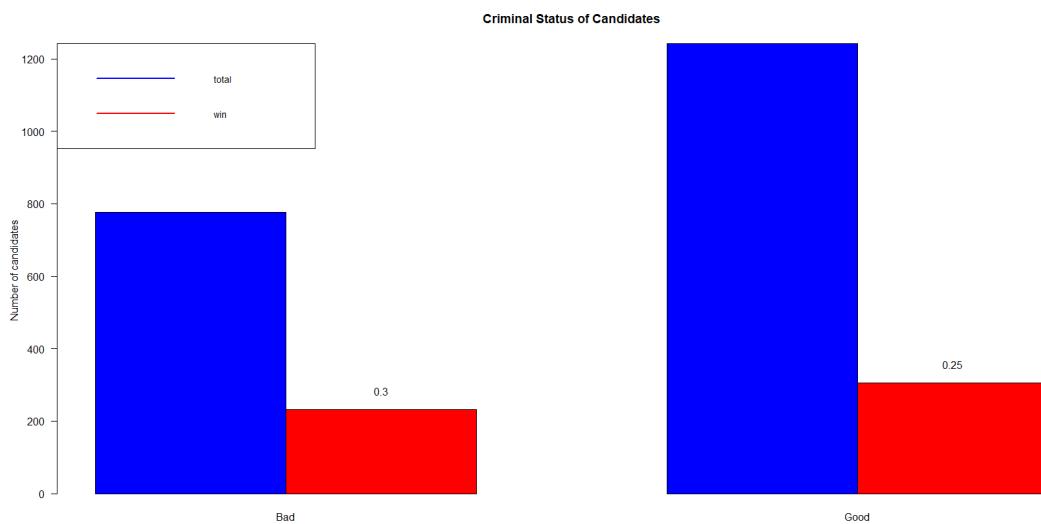
Category/Forward caste, suggesting that the majority of political participation was still caste related. Forward caste had a higher probability (0.287) of winning than SC and ST, which can also be seen from the higher red bar of General and the lower red bars of SC and ST than green bars.

## 2.7 Candidates by Gender

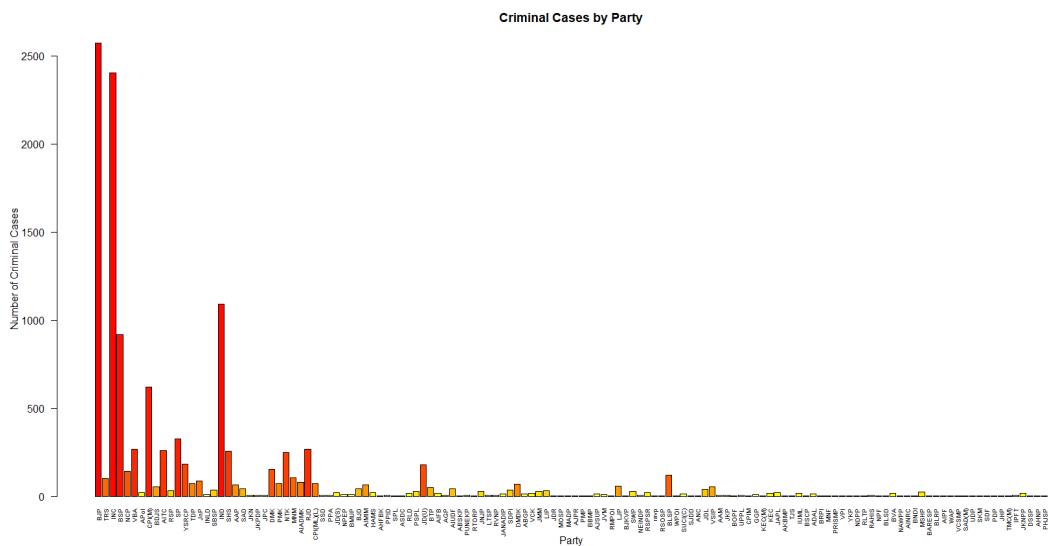


Among the total candidates, there were far more men than women. Also, there is a very clear gap between number of male and female representatives in the assembly. However, female had a slightly higher winning percentage (0.29).

## 2.8 Criminal Cases against Candidates



We classify candidates with no criminal case as Good, otherwise as Bad. Among the total candidates, a greater proportion had no criminal history. However, the winning candidates were almost evenly divided between the Good and Bad. Candidates with criminal cases even had a higher probability of winning (0.3)! It looks like having criminal cases made no difference to the outcome of the election.



BJP and INC gained the most seats in the 2019 Lok Sabha election, even though BJP Candidates had highest number of criminal cases against them and INC candidates were not too far behind.

### **3. Winner Prediction using Machine Learning Algorithms**

In order to predict the target variable "WINNER", we first split the dataset into training set (around 60%) and test set (around 40%). Since the number of winning candidates (539) were much less than losing candidates (1,479), this dataset is imbalanced. Therefore, we select 60% of the 539 winning candidates and 60% of the 1479 losing candidates to be our training data, and the remaining 40% of the 539 winning candidates and the remaining 40% of the 1,479 losing candidates become test dataset.

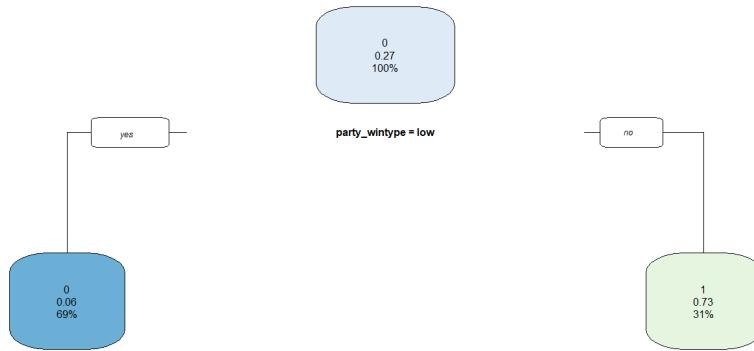
```
> idx_win <- which(df1$WINNER == 1)
> idx_lose <- which(df1$WINNER == 0)
> train_rate <- 0.6
> set.seed(1)
> train_idx <- c(sample(idx_win, size = floor(length(idx_win)*train_rate), replace = FALSE), sample(idx_lose, size = floor(length(idx_lose)*train_rate), replace = FALSE))
> df1_train <- df1[train_idx,]
> df1_test <- df1[-train_idx,]
```

Based on the previous visualization analysis, we select GENDER, CATEGORY, Criminal Status, Education Levels, Types of Parties as the variables for prediction,

### 3.1 Decision Tree

Decision tree is a classification model which works on the concept of information gain at every node. For all the data points, decision tree will try to classify data points at each of the nodes and check for information gain at each node. It will then classify at the node where information gain is maximum. It will follow this process subsequently until all the nodes are exhausted or there is no further information gain. Though Decision tree model is simple and of low predictive power, but the variables nodes pick are good hints for interpretation.

```
> fit <- rpart(WINNER~., data = df1_fac, method = 'class')
> rpart.plot(fit, extra = 106)
```



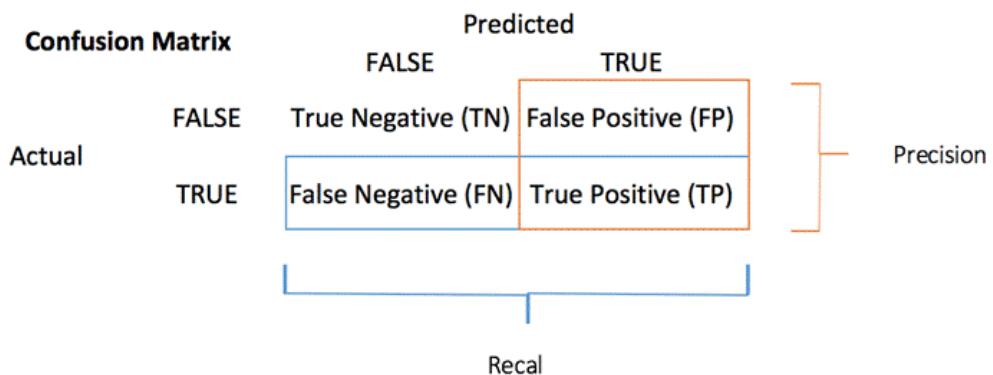
We start at the root node (the top of the graph): At the top, it is the overall probability of winning. It shows the proportion of candidates that won the election, 27 percent of total candidates. Then this node asks whether the party of the candidate is of low winning rate type. If yes, then the chance of winning is only 6%. rpart() function uses the Gini impurity measure to split the note. The higher the Gini coefficient, the more different instances within the node.

Next we can predict our test dataset, and create a table to count how many candidates are classified as winners and lost compared to the correct classification.

```
> predict_unseen <- predict(fit, df1_fact, type = 'class')
> table_mat <- table(df1_fact$WINNER, predict_unseen)
> table_mat
predict_unseen
  0   1
0 515 77
1 41 175
```

The model correctly predicts 515 losing candidates but classified 41 winners as losers. By analogy, the model misclassifies 77 candidates as winners while they turned out to be losers.

To measure the model performance, we can compute an accuracy measure for classification task with the confusion matrix. The general idea is to count the number of times True instances are classified as False.



Each row in a confusion matrix represents an actual target, while each column represents a predicted target. The first row of this matrix considers losing candidates (the False class): 515 were correctly classified as losers (True negative), while the remaining 77 was wrongly classified as winners (False positive). The second row considers the winners, the positive class were 175 (True positive), while the False negative was 41.

The accuracy test from the confusion matrix is the proportion of true positive and true negative over the sum of the matrix.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

```
> accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
> accuracy_Test
[1] 0.8539604
```

We have a score of 85 percent for the test set.

### 3.2 Random Forest

Random Forest is one such very powerful ensembling machine learning algorithm which works by creating multiple decision trees and then combining the output generated by each of the decision trees. Random Forest works on the same principle as Decision Tress; however, it does not select all the data points and variables in each of the trees. It randomly samples data points and variables in each of the tree that it creates and then combines the output at the end. It removes the bias that a decision tree model might introduce in the system. Also, it improves the predictive power significantly.

We construct 500 decision trees for this random forest to achieve good performance. In addition, for each bagged tree in this forest, only two variables are randomly sampled as candidates at each split.

```
> rf <- randomForest(WINNER ~ . , data=df1_fac, ntree=500,
+                         mtry=2, importance=TRUE)

> rf

call:
randomForest(formula = WINNER ~ ., data = df1_fac, ntree = 500,      mtry = 2, importance = TRUE)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2

      OOB estimate of  error rate: 12.73%
confusion matrix:
  0   1 class.error
0 785 102  0.1149944
1  52 271  0.1609907
```

Error rate is 12.73%. Then we use the model to predict on the test dataset and calculate misclassification error rate.

```
> df1_fact <- df1_test %>% mutate_if(is.character, as.factor)
> df1_fact$WINNER <- as.factor(df1_fact$WINNER)
> df1_fact <- df1_fact[,c(4,10,20,21,22)]
> prediction <- predict(rf, newdata=df1_fact, type="class")
```

```

> table(prediction, df1_fact$WINNER)
prediction  0   1
      0   516 43
      1    76 173
> misclassification_error_rate <- sum(df1_fact$WINNER != prediction) /
+  nrow(df1_fact)*100
> misclassification_error_rate
[1] 14.72772

```

The misclassification rate on the test dataset is 14.7%. Finally, we got the list of variable importance measures. Party type is the most important feature.

```

> importance(rf)
          0           1 MeanDecreaseAccuracy MeanDecreaseGini
CATEGORY      0.9880136 -6.429692      -5.628332      5.129235
party_wintype 148.2875485 158.642633     162.781975     223.384412
edu_level     -6.5485540  3.007133     -1.111387      8.116682
crim_class     0.4978169 -5.008002     -5.070179      3.495069

```

We achieved an accuracy of about 85.3% using Random Forest Classifier.

### 3.3 Logistic Regression

```

> glm1 <- glm(WINNER ~ crim_class + edu_level + party_wintype + GENDER + CATEG
ORY, data = df1_train, family = binomial)

> summary(glm1)

Call:
glm(formula = WINNER ~ crim_class + edu_level + party_wintype +
    GENDER + CATEGORY, family = binomial, data = df1_train)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-2.3695 -0.3830 -0.3405  0.3529  2.7400 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.77589  0.58208  4.769 1.85e-06 ***
crim_classGood -0.03704  0.19202 -0.193 0.847020  
edu_levelHigh_edu -0.05494  0.22153 -0.248 0.804140  
edu_levelLow_edu -0.87653  0.43819 -2.000 0.045463 *  
edu_levelOthers_edu  0.33503  0.50355  0.665 0.505841  
party_wintypelow -5.28349  0.53986 -9.787 < 2e-16 ***
party_wintypemed -1.83229  0.53425 -3.430 0.000604 *** 
GENDERMALE     -0.03100  0.26856 -0.115 0.908102  
CATEGORYSC     -0.27785  0.25264 -1.100 0.271433  
CATEGORYST     -0.24310  0.28737 -0.846 0.397579  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1404.07  on 1209  degrees of freedom
Residual deviance: 800.96  on 1200  degrees of freedom
AIC: 820.96

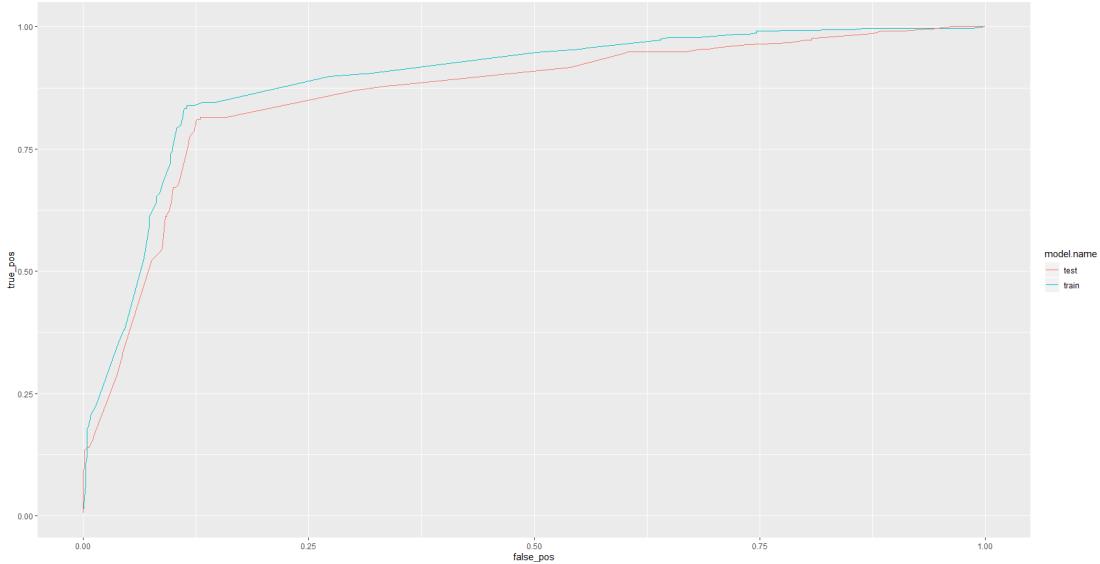
Number of Fisher scoring iterations: 5

```

Types of parties and low education level have small p-value, meaning they are significant in explaining winning possibility. We then use the model to predict winning rate in the test dataset, and

calculate the Receiver Operating Characteristics (ROC) curve from predicted probabilities and the ground truth for both training and testing data.

```
> prob <- predict(glm1, newdata = df1_test, type = "response")
```



```
> ROCs %>% group_by(model.name) %>%
+   mutate(delta=false_pos-lag(false_pos)) %>%
+   summarize(AUC=sum(delta*true_pos, na.rm=T)) %>%
+   arrange(desc(AUC))

# A tibble: 2 x 2
  model.name     AUC
  <fct>      <dbl>
1 train       0.905
2 test        0.876
```

Test error of the model obtained: AUC = 0.876 for test data set. The training and the test sets have very similar and high AUC, the model can make well predictions.

## 4. Conclusion

In this report, we conduct exploratory data analysis and winning rate prediction using Machine Learning algorithms on the 2019 Lok Sabha election data. Some stylized facts are forward caste and highly educated people made up the majority of the candidates; the likelihood of winning was independent of age or gender; only a few parties won the majority seats and criminal cases had no adverse effect on the election result.

We applied the methods of Decision Tree, Random Forest, and Logistic Regression to the train data and make prediction on the test data. Logistic regression achieved the highest accuracy of about 87.6%.