

|  |
| --- |
| Project – Machine Learning |
|  |
| **July 18, 2022**  PGP DSBA JAN 22A,  Great Learning  Authored by: Sharjil Shah |

|  |
| --- |
| Contents [Contents 2](#_Toc109095142)  [1.1 Read the dataset. Do the descriptive statistics and do the null value condition check and write an inference on it. 6](#_Toc109095143)  [a) Descriptive statistics of the dataset 6](#_Toc109095144)  [b) Checking for null values: 7](#_Toc109095145)  [c) Skewness of the dataset 7](#_Toc109095146)  [d) Inference: 7](#_Toc109095147)  [1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. 8](#_Toc109095148)  [a) Exploratory Data Analysis: 8](#_Toc109095149)  [i) Inferences: 9](#_Toc109095150)  [b) Univariate Analysis: 10](#_Toc109095151)  [c) Bivariate Analysis 12](#_Toc109095152)  [1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). 17](#_Toc109095153)  [a) Data Encoding: 17](#_Toc109095154)  [b) Scaling: 17](#_Toc109095155)  [c) Train and Test Split: 17](#_Toc109095156)  [1.4 Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models. 18](#_Toc109095157)  [a) Apply Logistic Regression: 18](#_Toc109095158)  [b) Apply Linear Discriminant Analysis: 19](#_Toc109095159)  [1.5 Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model. 21](#_Toc109095160)  [a) Apply KNN model: 21](#_Toc109095161)  [b) Apply Gaussian Naïve Bayes model: 22](#_Toc109095162)  [c) Comparison of both KNN and Naive Bayes model: 23](#_Toc109095163)  [1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. 23](#_Toc109095164)  [a) Tune the model using GridSerachCV and apply to Logistic Regression: 23](#_Toc109095165)  [b) Boosting = Ada Boosting 26](#_Toc109095166)  [1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. 27](#_Toc109095167)  [a) Logistic Regression 27](#_Toc109095168)  [b) Linear Discrimination Analysis: 29](#_Toc109095169)  [c) KNN: 30](#_Toc109095170)  [d) Naive Bayes 32](#_Toc109095171)  [e) Ada Boost 33](#_Toc109095172)  [f) Gradient Boost 35](#_Toc109095173)  [g) Bagging using Random Forest 36](#_Toc109095174)  [h) Comparison of Performance Metrics 38](#_Toc109095175)  [1.8 Based on these predictions, what are the insights. 38](#_Toc109095176)  [2.1 Find the number of characters, words, and sentences for the mentioned documents. 39](#_Toc109095177)  [2.2 Remove all the stopwords from all three speeches. 39](#_Toc109095178)  [2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (After removing the stopwords) 39](#_Toc109095179)  [2.4 Plot the word cloud of each of the three speeches. (after removing the stopwords) 40](#_Toc109095180)  [A) Roosevelt: 40](#_Toc109095181)  [B) Kennedy: 41](#_Toc109095182)  [C) Nixon: 41](#_Toc109095183) |
|  |
|  |

Problem 1:

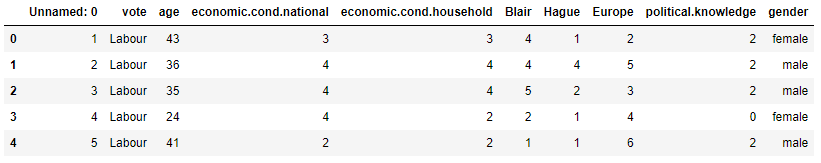
You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Dataset for Problem: [Election\_Data.xlsx](https://olympus.mygreatlearning.com/courses/67081/files/5908264/download?verifier=wO8tbnNng2GV5K8ZHBY4WSbDUS1CCM46pxLBq7j1&wrap=1)

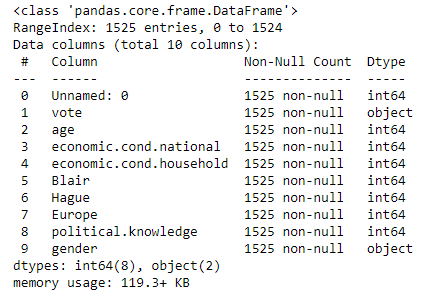
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| vote | Party choice: Conservative or Labour |
| age | in years |
| economic.cond.national | Assessment of current national economic conditions, 1 to 5 (1 – poor, 5- great) |
| economic.cond.househol d | Assessment of current household economic conditions, 1 to 5 (1 – poor, 5- great) |
| Blair | Assessment of the Labour leader, 1 to 5 |
| Hague | Assessment of the Conservative leader, 1 to 5 |
| Europe | Assessment of the Conservative leader, 1 to 5 |
| political.knowledge | Knowledge of parties&apos; positions on European integration, 0 to 3 |
| gender | female or male |

# Read the dataset. Do the descriptive statistics and do the null value condition check and write an inference on it.

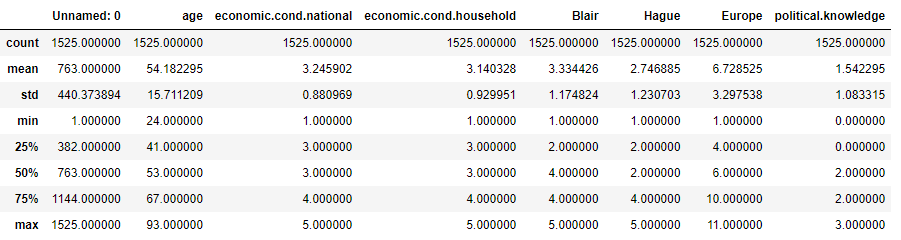
Let’s take a look at the head and tail of the dataset.



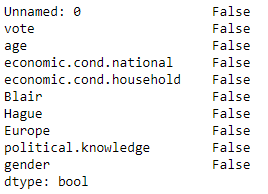
The dataset has no null values, there are 8 variables that are integers and 2 variables are objects.

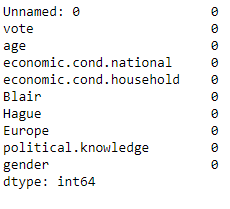


## Descriptive statistics of the dataset

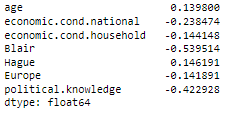


## Checking for null values:





## Skewness of the dataset



## Inference:

* With the problem statement we know that the target variable is ‘Vote’ from the dataset.
* The head and tail of the dataset tells us that there are 2 main parties for whom the voters are voting for and they are namely: ‘Labour’ and ‘Conservative’.
* The dataset has 10 unique columns, out of which 2 are objects and 8 are integers. Column “Unnamed:0” is an index column and will be dropped while performing EDA.
* From the descriptive statistics we can see that the youngest voter is of the age 24, 50% of the voters are of the age 53 and the oldest voter is 93 years old
* Labour party seems to be bagging more number of votes and the most number of voters are females
* Variables ‘economic.cond.national’, ‘economic.cond.household’, ‘Blair’, ‘Hague’, ‘Europe’ and ‘political.knowledge’ are ordinal variables.
* 50% of the voters have assessed ‘Blair’ who is the leader of Labour Party to be at 4 which is higher than that of ‘Hague’ who is the leader of Conservative Party

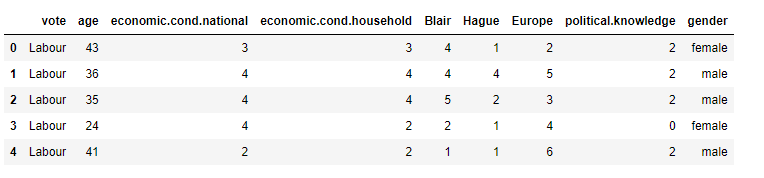
# Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

## Exploratory Data Analysis:

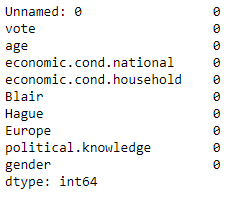
The dataset has 10 columns and 1525 rows.



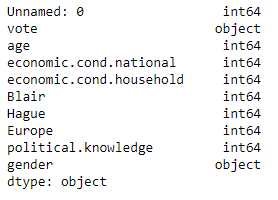
Column “Unnamed: 0” should be dropped since it does not have any significance in this study Snippet below shows the head and tail after dropping column “Unnamed: 0”



Although we know from section 1.1 that there are no null values, the snippet below proves the point further. (Note: Please check the code for this inference from jupyter notebook)



Similarly, let’s also check the data types of each variable. Variable “vote” and “gender” are objects and the rest are integers.



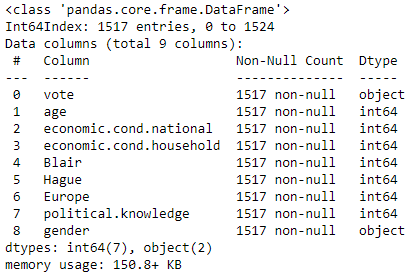
The dataset has 8 duplicate records which will be dropped as they do not add any value to the study



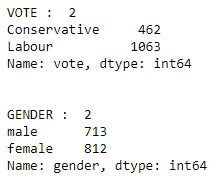
Let’s take a look at the shape of the dataset after dropping the duplicated records and column “Unnamed: 0”



As seen in section 1.1, variables “economic.cond.national”, “economic.cond.household”, “Blair”, “Hague”, “Europe” and “political.knowledge” are ordinal variables and must be converted to object data type. Info of the variables after converting the variables.



Let’s check the unique values in the categorical/ object variables



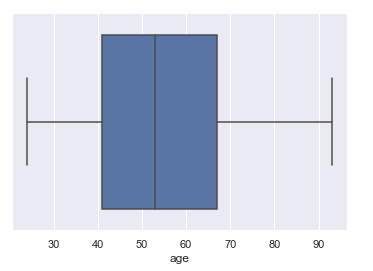
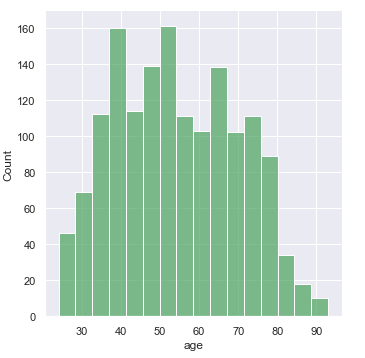
Since variable “vote” is our target variable and it has 2 categories.

### Inferences:

* By doing initial EDA, we can say that the original data set had 1525 rows and 10 columns
* There are 2 variables whose datatype is “object” and these are categorical variables
* The other variables “economic.cond.national”, “economic.cond.household”, “Blair”, “Hague”, “Europe” and “political.knowledge” are also ordinal/ categorical and hence their data type has been converted
* The dataset does not have any null or missing values and none of the categorical variables have a as “?” or “blank”
* There were 8 duplicate records which were dropped as they do not add any value to our analysis
* The percentage of votes are not balanced between the two parties where 69.68% of the voters voted for Labour party and only 30.32% of the voters voted for Conservative party.

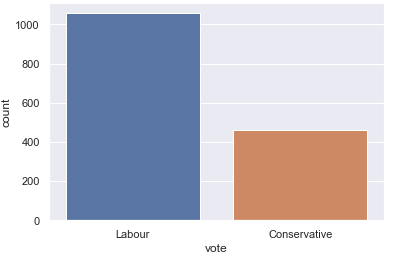
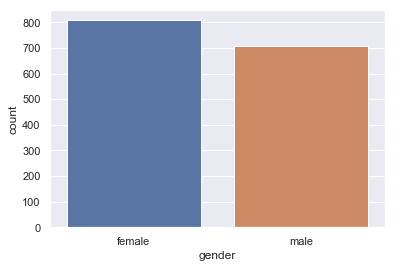
## Univariate Analysis:

1. Figure below shows the Univariate Analysis using Distplot and Boxplot of variable “age”



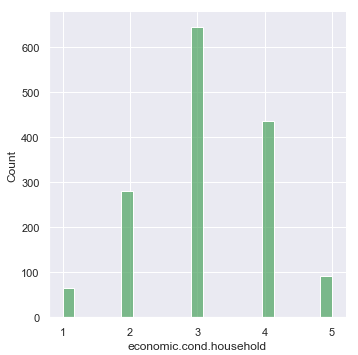
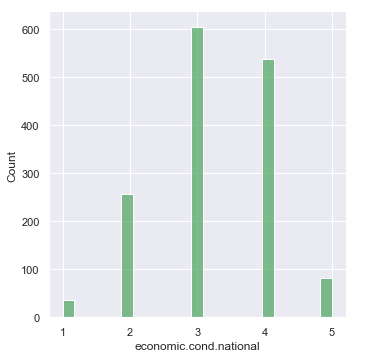
There is a normal distribution in variable “age”. Most of the voters are found to be between the age of 40 to 80.

* + 1. Figure below shows the Univariate Analysis using Countplot of variables “vote” and “gender”



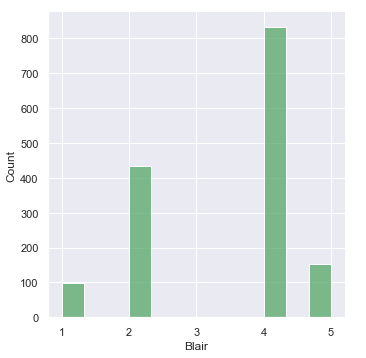
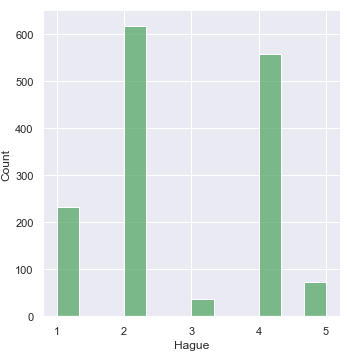
We can infer that the Labour party is being favored more by the voters. The ratio of female to male is almost the same with female voters being more than male

* + 1. Figure below shows the Univariate Analysis using Countplot of variables “economic.cond.national” and “economic.cond.household”



Variable “economic.cond.national” and “economic.cond.household” show that most of the voters have rated these two scales as 3 and 4 which is a moderate rating.

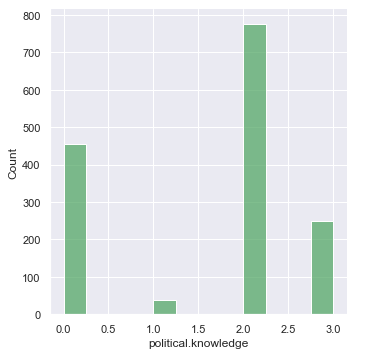
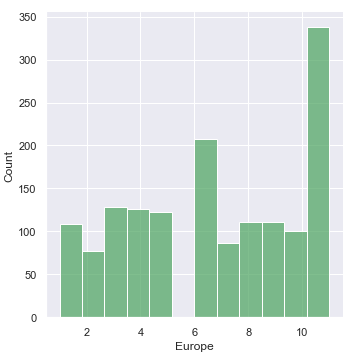
* + 1. Figure below shows the Univariate Analysis using Countplot of variables “Blair” and “Hague”



Most of the voters have voted “Blair” as 4 compared to “Hague”

Many voters have rated “Hague” as 2 compared to “Blair”

* + 1. Figure below shows the Univariate Analysis using Countplot of variables “Europe” and “political.knowledge”



In a scale of 1 to 11, most of the voters have voted that the European integration is between 2 to 10

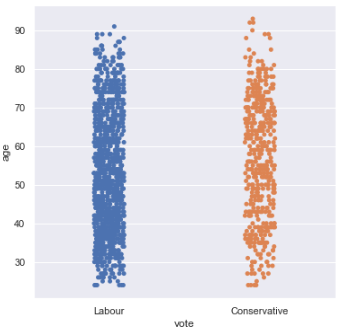
with maximum as 11. Hence, most of them have an inclination that the parties represent

‘Eurosceptic’ sentiment.

In a scale of 0 to 3, most of the voters have voted the Political knowledge to be 2 which is moderate.

## Bivariate Analysis

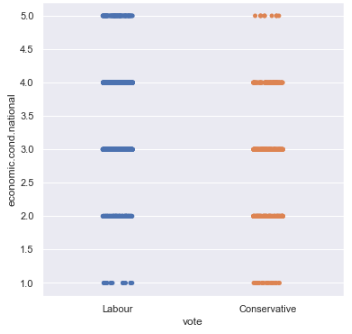
* + 1. Figure below shows the Bivariate Analysis using Strip plot which are taken from jupyter notebook



Nearly similar trend. Since there are more voters for Labour party, the strip looks denser for Labour

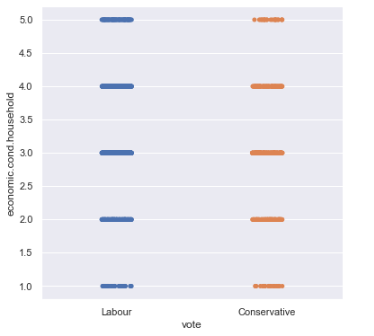
party. One key difference would be that most of the voters above the age of 90 have voted for Conservative

party.



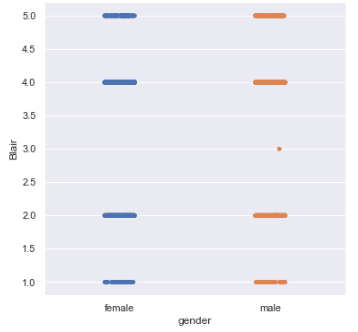
We can see that most of the voters have rated Labour party as 5 compared to Conservative Party for their

assessment on current national economic conditions. Otherwise the trend is nearly the same.



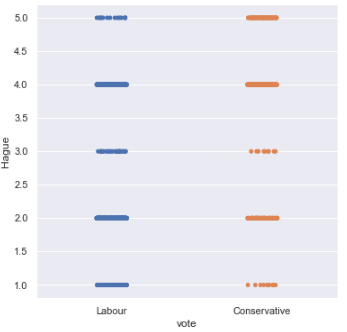
Trend is nearly the same. Except that more voters have rated Labour party to be a 5 when it comes to an

assessment on economic household conditions.

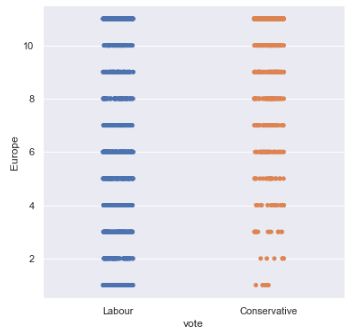


Trend is the same with one exception that none of the female voters rated Blair as 3 but few male voters

have rated him as 3.

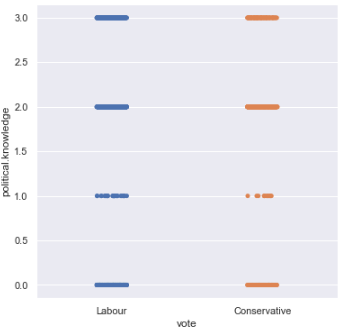


Trend is nearly the same.



Many voters have rated Labour party between the scale of 2 to 5 compared to the Conservative Party.

Otherwise the trend is the same.

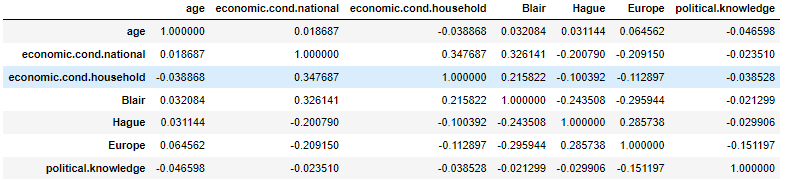


On a scale of 0 to 3, Labour party seems to have more voters rating them as 1 compared to Conservative

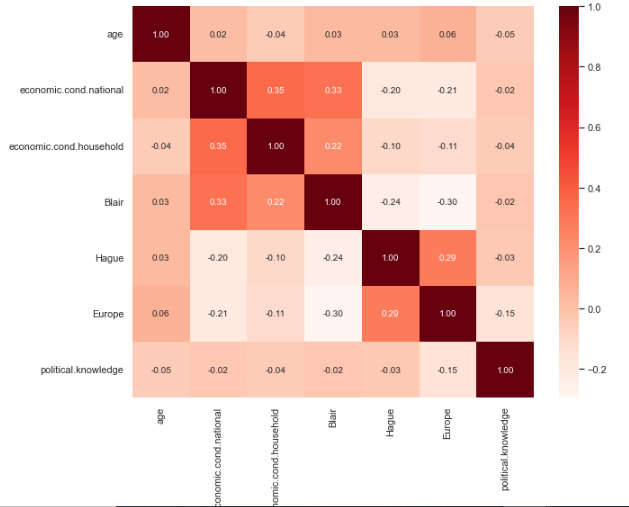
Party. Otherwise the trend is the same.

* + 1. Figure below shows the Correlation Matrix and Heat Map which are taken from the jupyter notebook

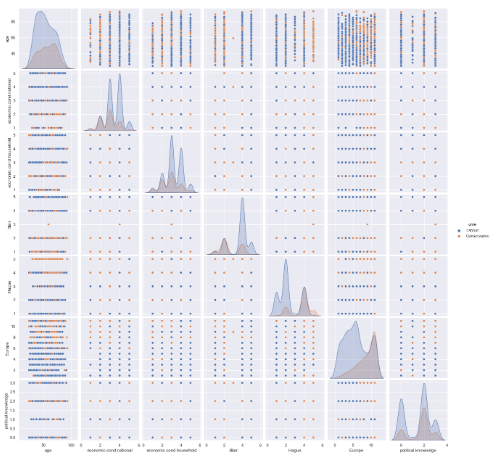
Correlation Matrix before converting all the variables except “age” to categorical variables



Heat Map before converting all the variables except “age” to categorical variables

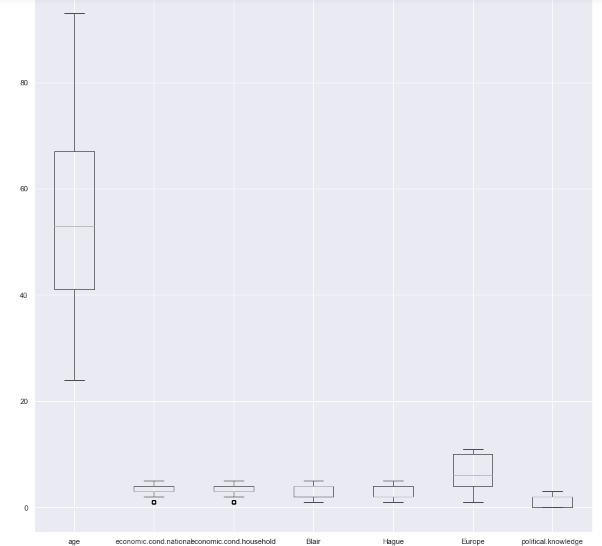


* The heat map shows that there is no high correlation between any of the variables as most of the values are under 0.35
* There is negative correlation between age and & political knowledge and “economic.cond.household”
* Variables “economic.cond.national” and “economic.cond.household” have the highest correlation of 0.35
  + 1. Figure below shows the Pair Plot which is taken from the jupyter notebook.



No correlation is found between the variables

* + 1. Figure below shows if there are outliers present which is taken from the jupyter notebook



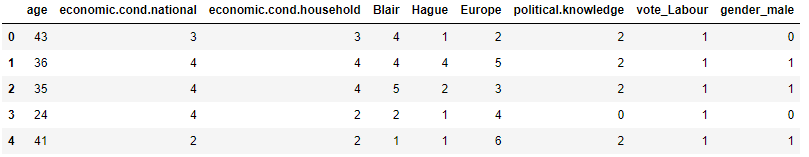
* No outlier found in variable “age”
* Some outliers are found in variables “economic.cond.household” and “economic.cond.household” which can be checked in Univariate analysis done in jupyter notebook. Since those are ordinal variables, we will not be treating them.

# Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

## Data Encoding:

* Variable “vote” & “gender” contain string values. To perform the analysis for this dataset, we would be converting the string values to the integer data type as modelling cannot take string / object data types.
* For converting variable to numerical data we will use pd.Categorical().codes function
* Variable “gender” has ‘male’ and ‘female’ which will be converted to 1 and 0 respectively.
* Variable “vote” has “Labour” and “Conservative” which will be converted to 1 and 0 respectively

Snippet from jupypter notebook shows that that column “gender” has been changed to numerical data



Copy all the predictor variables into X dataframe

Copy target into the y dataframe.

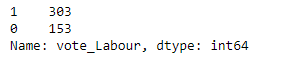
## Scaling:

* Scaling is done on continuous variables in a dataset with different unit of measures.
* All variables are either categorical or ordinal except for variable “age”.
* For Logistic regression, LDA and Naïve Baye’s model we need not perform any scaling, however, for KNN it is necessary to scale the data, as it a distance-based algorithm (typically based on Euclidean distance).
* For KNN, the variables will be scaled using the min max scaler.

## Train and Test Split:

The data will be split using train\_test\_split() with random state = 1 and test size = 0.30.





# Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models.

## Apply Logistic Regression:

The data has been pre-processed and has been split to train and test data with test size = 0.30 in the previous sections.

Step 1: We will apply Logistic Regression with below parameters.

Step 2: Check the model score by using model.score()





Step 3: Check confusion matrix

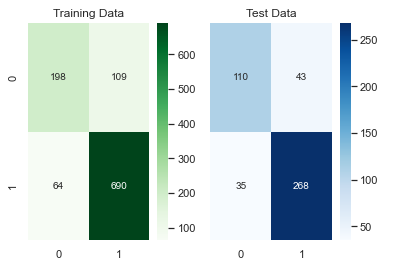
Train data:



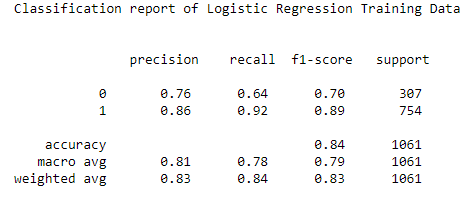
Test data:

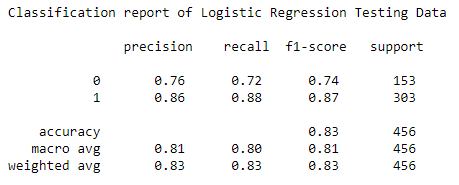


Heat Map of Confusion Matrix:



Step 4: Check classification matrix

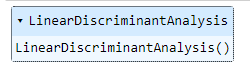




* Hyper parameters were added to see if the model behave differently in training and testing data. By using solver = liblinear, there was some differences found in the models.
* The accuracy of model in training set is 0.84 and on testing set is 0.83, which is good and very close to each other.
* The recall of Conservative party is better on Testing data whereas the recall of Labour party is better on Training data
* Overall it is a good model and there is no over fitting found.

## Apply Linear Discriminant Analysis:

Step 1: We will apply LDA to the training and testing data.



Step 2: Check the model score by using model.score()





Step 3: Check confusion matrix

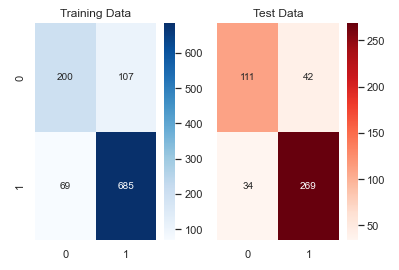
Train data:



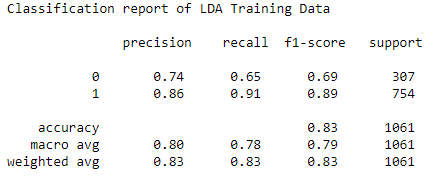
Test data:

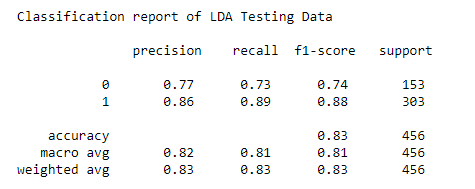


Heat Map of confusion matrix



Step 4: Check classification matrix





* The accuracy of model in training set and testing set is the same which is 0.83.
* The recall of Conservative party is better on Testing data whereas the recall of Labour party is better on Training data
* Overall the model is performing well.

Comparison of two models:

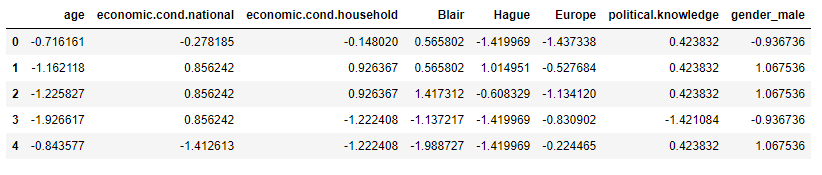
While comparing both these models, we find both results are almost same, but LDA works better

since the recall with LDA is slightly better on Testing data.

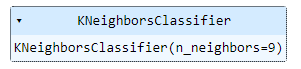
# Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model.

## Apply KNN model:

Step 1: Scale the data



Step 2: We will apply KNN with below parameters.



Steps 3: Check the model score by using model.score().





Step 4: Check confusion matrix:

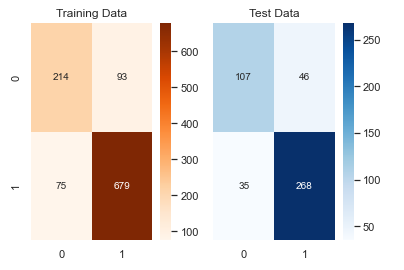
Train data:



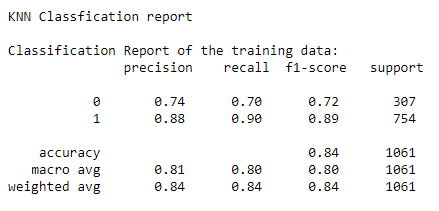
Test data:

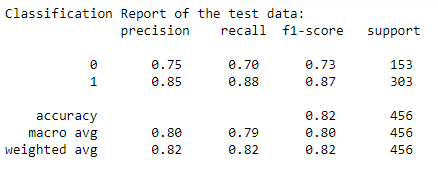


Heat Map of confusion matrix



Step 5: Check classification matrix

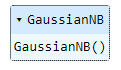




* The accuracy of model is 0.84 and 0.82 on training and testing data respectively
* The recall of Conservative party is better on Testing data and Labour is slightly better on Training data

## Apply Gaussian Naïve Bayes model:

Step 1: We will apply Gaussian Naïve Bayes with below parameters



Step 2: Check the model score by using model.score()





Step 3: Check confusion matrix

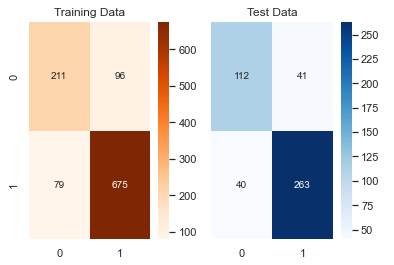
Train data:



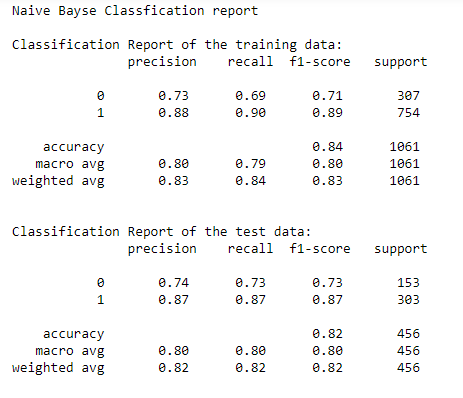
Test data:



Heat Map of confusion matrix



Step 4: Check classification matrix



* The accuracy of model in training set is 0.84 and on testing set is 0.82. Hence, the model is performs better on Training data.
* The recall of Conservative party is better on Testing data whereas the recall of Labour party is better on Training data

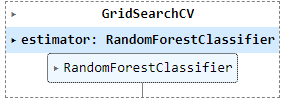
## Comparison of both KNN and Naive Bayes model:

* Both models are good and does not overfit or underfit. The accuracy of both models on Training data is the same, however, accuracy of KNN is better on testing data
* KNN model is better compared to Naïve Bayes after applying hyperparameters as the model has better recall.

# Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.

## Tune the model using GridSerachCV and apply to Logistic Regression:

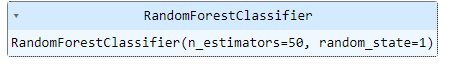
Step 1: Tune Logistics Regression model using GridSearchCV



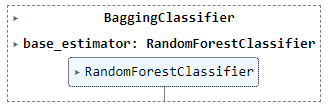
Step 2: Best parameters and estimators



Step 3: Random forest classifier



Step 4: bagging with random forest classfier



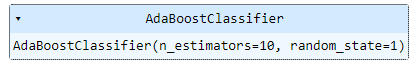
1. Bag score for train:



1. Bag score for test:



1. Boosting = Ada Boosting



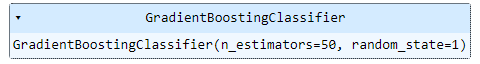
1. Train Score:



1. Test Score:



1. Boosting = Gradient Boosting



1. Train score:



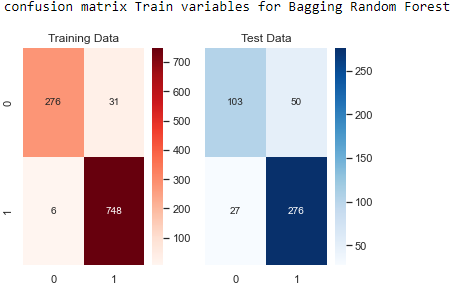
1. Test score:



1. Bagging random forest



1. Confusion matrix for Bagging Random Forest



1. Confusion matrices:

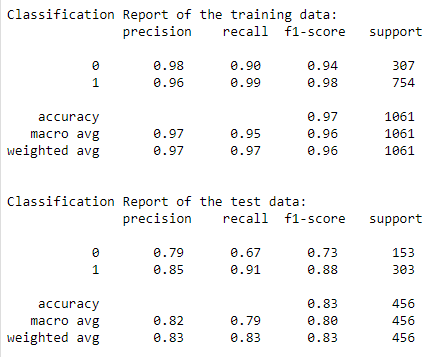
Train:



Test:



1. Bagging (Random Forest) Classfication report



## Boosting = Ada Boosting

1. Model score:



1. Confusion matrix Train variables for ADA Boosting:



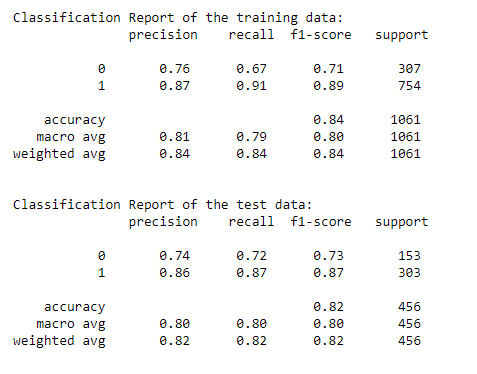
Train data:



Test data:



1. ADA Boosting Classfication report:



# Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

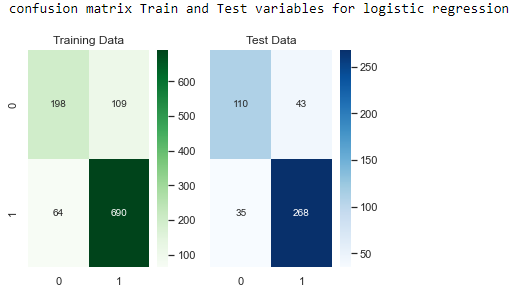
## Logistic Regression

1. Model Accuracy:



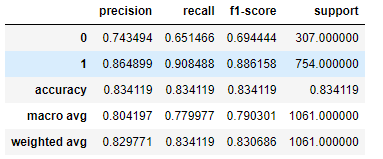


1. Heat Map of confusion matrix:

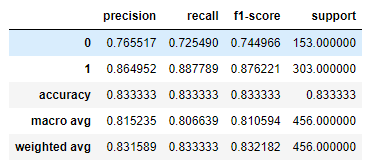


1. Classification matrix:

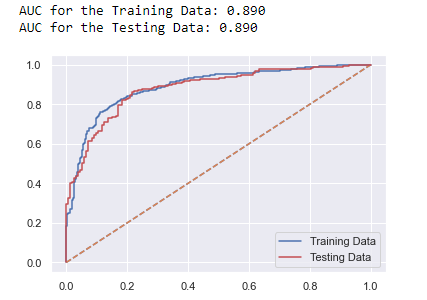
Train data:



Test data:



1. AUC and ROC



## Linear Discrimination Analysis:

1. Model Accuracy:



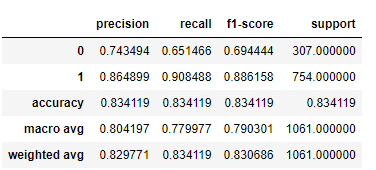


1. Heat Map of confusion matrix

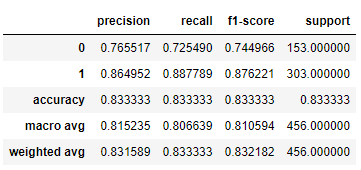


1. Classification matrix:

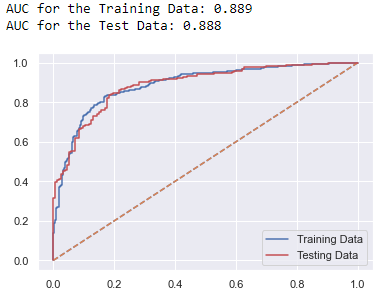
Train:



Test:



1. AUC and ROC

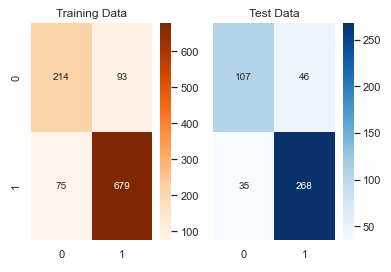


## KNN:

1. Model Accuracy

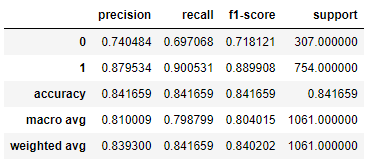


1. Heat Map of confusion matrix

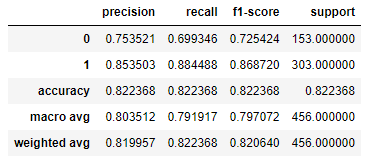


1. Classification matrix

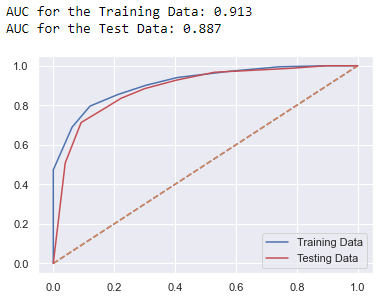
Train



Test:



1. AUC and ROC:



## Naive Bayes

* 1. Model Accuracy

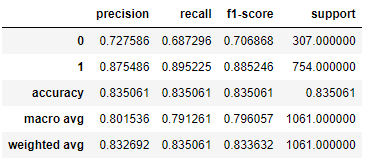


* 1. Heat Map of confusion matrix

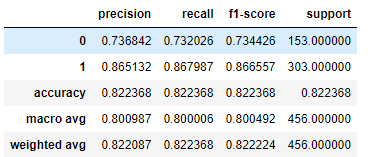


* 1. Classification matrix

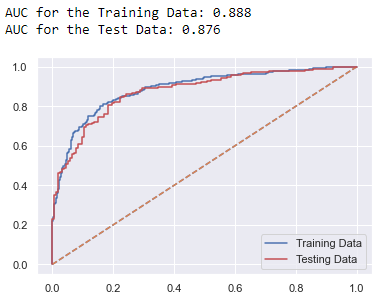
Training data:



Testing data:



1. AUC and ROC

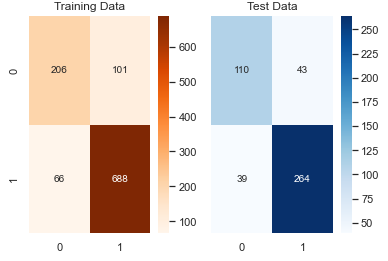


## Ada Boost

* 1. Model accuracy

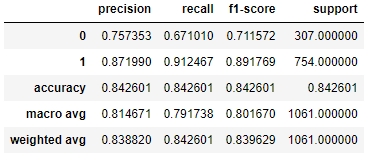


* 1. Heat Map of confusion matrix:

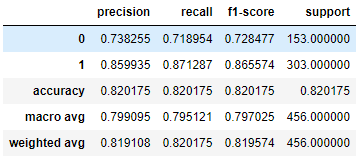


* 1. Classification matrix:

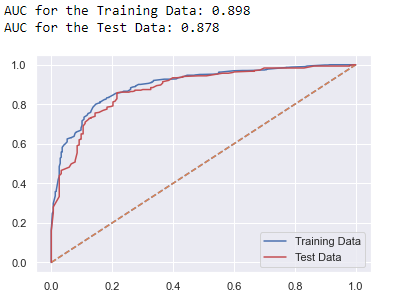
Training data:



Testing data:



* 1. AUC and ROC:

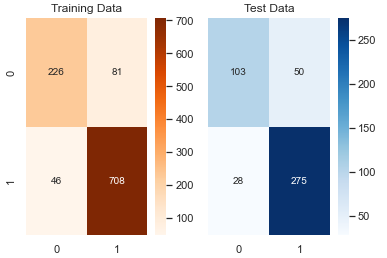


## Gradient Boost

1. Model accuracy

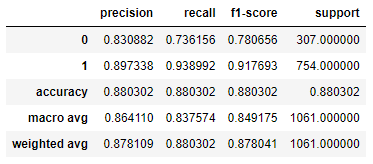


1. Heat Map of confusion matrix

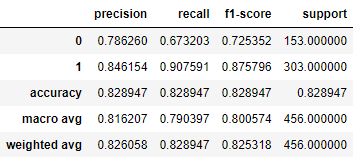


1. Classification matrix

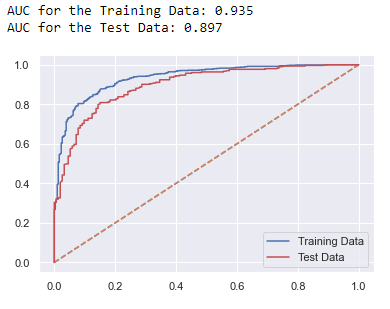
Training data:



Testing data:



1. AUC and ROC

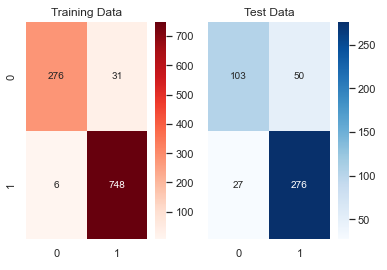


## Bagging using Random Forest

1. Model Accuracy

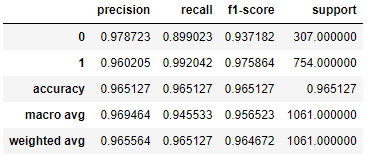


1. Heat Map of confusion matrix

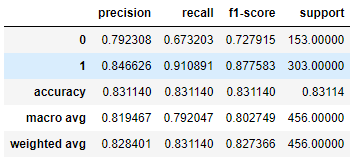


1. Classification matrix

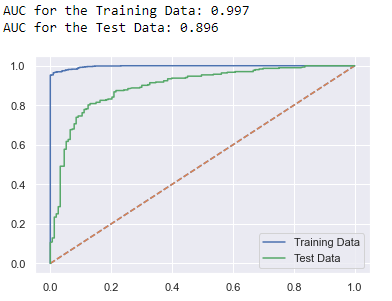
Training data:



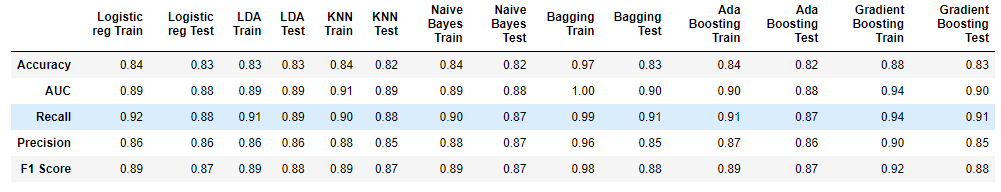
Testing data:



1. AUC and ROC



# Comparison of Performance Metrics



By comparing the performance metrics, we can conclude the following:

* Logistic Regression, LDA, KNN and Gausion Naïve Bayes are good models because they work well on both Training and Testing data with model accuracy similar across both training and testing data.
* However, LDA has better accuracy and recall and fi-score.
* Gradient Boosting and Bagging using Random Forest is not a good model because it is overfitting on training data and doesn’t perform well on testing data

# 1.8 Based on these predictions, what are the insights.

Comparing all the Models we see that Logistic Regression, LDA, KNN and Gradient Boosting are good models, however, LDA Model gives better results.

* We observe Labour has higher possibility of winning
* Labour has higher voting possibility among all age groups except for very old people.
* Irrespective of the political knowledge levels or gender, Labour has an edge on higher votes

• Where the Eurosceptic sentiment is more, Conservative has scope for winning

Problem 2

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

# 2.1 Find the number of characters, words, and sentences for the mentioned documents.

Number of words:

1. Number of words in President Franklin D. Roosevelt speech 1360
2. Number of words in President John F.Kennedy speech 1390
3. Number of words in President Richard Nixon speech 1819

Number of characters:

1. Number of characters in President Franklin D. Roosevelt speech 7571
2. Number of characters in President John F.Kennedy speech 7618
3. Number of characters in President Richard Nixon speech 9991

Number of sentences:

1. Number of sentences in President Franklin D. Roosevelt speech 67
2. Number of sentences in President John F.Kennedy speech 52
3. Number of sentences in President Richard Nixon speech 68

# 2.2 Remove all the stopwords from all three speeches.

1. 184 stop words were identified
2. Stop words count in Roosevelt's Speech is 730
3. Stop words count in Kennedy's Speech is 711
4. Stop words count in Nixon's Speech is 1017
5. Word count in Roosevelt's Speech after removing stop words is 604
6. Word count in Kennedy's Speech after removing stop words is 652

Word count in Nixon's Speech after removing stop words is 784

# 2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (After removing the stopwords)

In Roosevelt’s speech, below words occur more frequently

1. Nation – 11 times
2. Spirit – 9 times
3. Democracy – 9 times

In Kennedy’s speech, below words occur more frequently

World – 8 times

1. new – 7 times
2. pledge – 7 times

In Nixon’x speech, below words occur more frequently

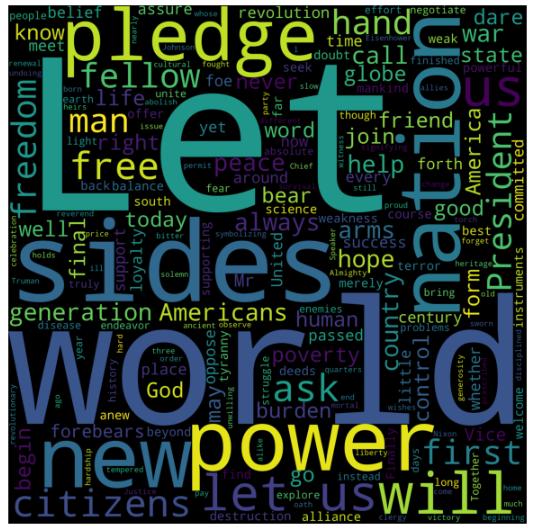
1. Peace – 19 times
2. World – 16 times
3. New – 15 times
4. America – 13 times

# 2.4 Plot the word cloud of each of the three speeches. (after removing the stopwords)

## Roosevelt:



## Kennedy:



## Nixon:

