**Machine Learning**

**Business Report**

Yedupati Venkata Yamini

28 May 2023

**Table of Contents**:

Problem 1:

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it…………………………………………………………………………………… 7

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers………………………………………………………………………………………… 12

Data Preparation:  
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)………………………………………………….. 22

Modelling:  
1.4 Apply Logistic Regression and LDA (linear discriminant analysis)………………………. 25

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results………………………… 27

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting… 29

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………………34  
1.8 Based on these predictions, what are the insights?................................................................. 62

Problem 2:

2.1 Find the number of characters, words, and sentences for the mentioned documents……. 65

2.2 Remove all the stopwords from all three speeches……………………………………… 66

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) ……………………………… 67

2.4 Plot the word cloud of each of the speeches of the variable.(after removing the stopwords) 68

**List of figures**:

Figure 1: Basic information of the dataset

Figure 2: Counts of null values for columns in the dataset

Figure 3: Box plot for Age

Figure 4: Information of the dataset after data pre-processing

Figure 5: Box plot for Age

Figure 6: Histogram for Age

Figure 7: Count plots for categorical variables

Figure 8: Value counts for all categorical variables

Figure 9: Pie charts for all categorical variables

Figure 10: Age distribution of voters voted for each party

Figure 11: Count plots of all categorical variables distinguished by party voted for

Figure 12: Age distribution of people with respect to assessment scores given to leaders

Figure 13: Percentages of voters voted to different parties w.r.t Gender and all other categorical variables

Figure 14: Basic information of encoded dataset

Figure 15: Predicted probabilities on training and test data after applying Logistic Regression

Figure 16: Feature importances from Random Forest Classifier

Figure 17: Feature importances from Adaptive Boosting Classifier

Figure 18: Feature importances from Gradient Boost classifier

Figure 19: Predicted probabilities – Logistic Regression

Figure 20: Confusion matrix – train and test data (Logistic Regression)

Figure 21: Classification report – training and test data (Logistic Regression)

Figure 22: AUC ROC curve – training data (Logistic Regression)

Figure 23: AUC ROC curve – testing data (Logistic Regression)

Figure 24: Predicted probabilities – Gaussian Naive Bayes

Figure 25: Confusion matrix – train and test data (Gaussian Naive Bayes)

Figure 26: Classification report – training and test data (Gaussian Naive Bayes)

Figure 27: AUC ROC curve – training data (Gaussian Naive Bayes)

Figure 28: AUC ROC curve – testing data (Gaussian Naive Bayes)

Figure 29: Predicted probabilities – K Nearest Neighbors

Figure 30: Confusion matrix – train and test data (K Nearest Neighbors)

Figure 31: Classification report – training and test data (K Nearest Neighbors)

Figure 32: AUC ROC curve – training data (K Nearest Neighbors)

Figure 33: AUC ROC curve – testing data (K Nearest Neighbors)

Figure 34: Predicted probabilities – Bagging Classifier

Figure 35: Confusion matrix – train and test data (Bagging Classifier)

Figure 36: Classification report – training and test data (Bagging Classifier)

Figure 37: AUC ROC curve – training data (Bagging Classifier)

Figure 38: AUC ROC curve – testing data (Bagging Classifier)

Figure 39: Predicted probabilities – Random Forest Classifier

Figure 40: Confusion matrix – train and test data (Random Forest Classifier)

Figure 41: Classification report – training and test data (Random Forest Classifier)

Figure 42: AUC ROC curve – training data (Random Forest Classifier)

Figure 43: AUC ROC curve – testing data (Random Forest Classifier)

Figure 44: Predicted probabilities – Adaptive Boosting

Figure 45: Confusion matrix – train and test data (Adaptive Boosting)

Figure 46: Classification report – training and test data (Adaptive Boosting)

Figure 47: AUC ROC curve – training data (Adaptive Boosting)

Figure 48: AUC ROC curve – testing data (Adaptive Boosting)

Figure 49: Predicted probabilities – Gradient Boosting

Figure 50: Confusion matrix – train and test data (Gradient Boosting)

Figure 51: Classification report – training and test data (Gradient Boosting)

Figure 52: AUC ROC curve – training data (Gradient Boosting)

Figure 53: AUC ROC curve – testing data (Gradient Boosting)

Figure 54: Feature importances of independent variables

Figure 55: Word cloud for Roosevelt’s speech

Figure 56: Word cloud for Kennedy’s speech

Figure 57: Word cloud for Nixon’s speech

**List of tables**:

Table 1: Top 5 rows of initial dataset

Table 2: Bottom 5 rows of initial dataset

Table 3: Top5 rows of dataset after removing Unnamed column

Table 4: Summary of the dataset

Table 5: Top 5 rows of encoded dataset

Table 6: Top 5 rows of independent variables of training dataset

Table 7: Top 5 rows of dependent variables of training dataset

Table 8: Top 5 rows of independent variables of testing dataset

Table 9: Top 5 rows of dependent variables of testing dataset

Table 10: Confusion matrix

Table 11: Performance metrics of all models on elections dataset

Table 12: Most frequently used words in Roosevelt’s speech

Table 13: Most frequently used words in Kennedy’s speech

Table 14: Most frequently used words in Nixon’s speech

**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/87097/files/8138804/download?verifier=CGu44yb3X3sOTOswXQzIQEOQ8SkF5SQLqHa50fzO&wrap=1)

Data Ingestion:

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

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

Data Preparation:  
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).

Modelling:  
1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

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

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

**1.1 Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.**

Ans:

Brief description of all the features in the dataset:

1. vote: Party choice: Conservative or Labour by the voter.
2. age: Age of the voter (in years)
3. economic.cond.national: Assessment of current national economic conditions, 1 to 5. 1 - poor economic conditions. 5 - high economic conditions
4. economic.cond.household: Assessment of current household economic conditions, 1 to 5. 1- poor household conditions, 5 - rich household economic conditions
5. Blair: Name of person contesting from Labour party. Assessment of the Labour leader, 1 to 5. score given by voterto Blair 1 - low 5 - high
6. Hague: Name of person contesting from Conservative party. Assessment of the Conservative leader, 1 to 5. score given by voter to Hague. 1 - low, 5 - high
7. Europe: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment. It corresponds to the voter.
8. political.knowledge: Knowledge of parties' positions on European integration by the voter, 0 to 3.
9. gender:Male or female voter.

Viewing the basic information of the dataset:

* The given data set has 1525 rows and 10 columns including dependent and independent variables.
* Viewing the column and data types information of the features in the dataset:

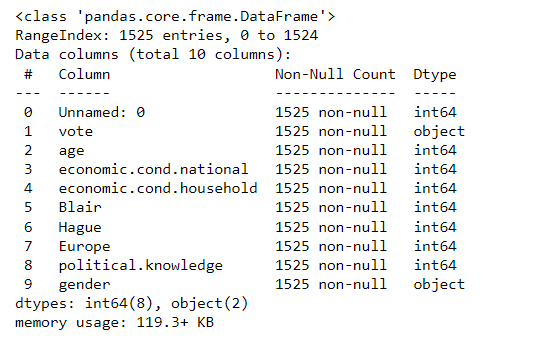


Figure 1: Basic information of the dataset

Insights:

1. The columns present in the dataset in the initial glance contain 2 categorical and 8 integer variables.

2. The column ‘Unnamed: 0’ seems to be the serial number of the rows in the dataset and it does not provide any value to the analysis. Hence, it can be dropped.

3. ‘Vote’ is the important feature variable which indicates the party that the voter has voted for.

4. Variables namely ‘economic.cond.national’, ‘economic.cond.household’, ‘Blair’, ‘Hague’, ‘Europe’, ‘political.knowledge’ are ordinal variables which indicate the level of magnitude for each feature.

5. There are no null values in any of the features. Hence, no imputation is needed.

Viewing the first 5 rows of the dataset:

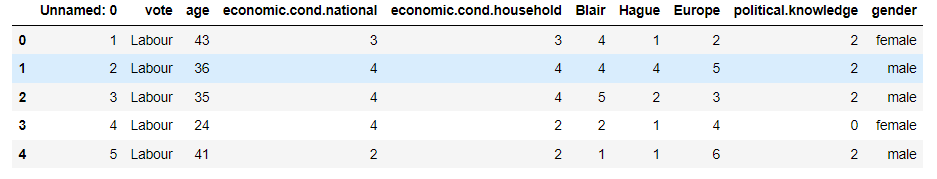


Table 1: Top 5 rows of initial dataset

Viewing the last 5 rows of the dataset:

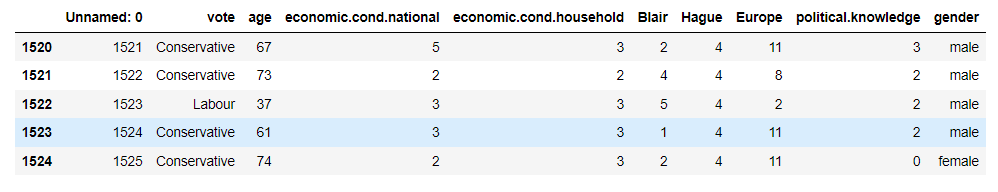


Table 2: Bottom 5 rows of initial dataset

We do not need the ‘Unnamed: 0’ column for our further analysis. The resultant dataset for further analysis after removing the column looks as follows:

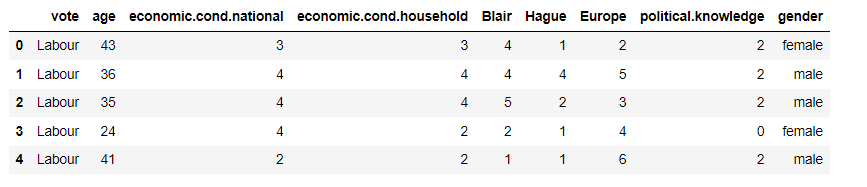


Table 3: Top5 rows of dataset after removing Unnamed column

Viewing the summary of the dataset:

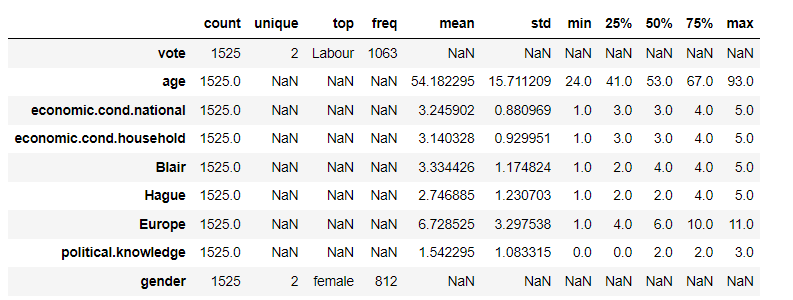


Table 4: Summary of the dataset

Insights:

1. Voters in the dataset have voted for 2 unique parties ‘Conservative’ and ‘Labor’. 1063 of the 1525 voters have voted for ‘Labor’ which seems to be a huge majority winning state for ‘Labor’ party.

2. The average age of the voters is around 54 years with 50% of the people between 41 and 67 years. The oldest voter in the region is 93 years old.

3. Almost half of the voters have assessed the current national economic condition to be below 3. It indicates their assessment of the national economic condition to be weak to moderate. Only 25% of the people have assessed the current national economic condition to be between 4 and 5.

4. Household economic conditions of more than half of the voters are weak to moderate. (below or same as 3)

5. At Least 50% of the voters have assessed Blair with a score of 4 whereas more than 50% of the voters have assessed Hague with a score of 2. It indicates that Blair from the Labor party has received higher assessment scores than Hague from Conservative party.

6. At Least 25% of the voters have no political knowledge.

7. There are more female voters than male voters but the difference between the number of male and female voters is not huge to be bothered.

8. At Least half of the voters have a highly tentative attitude towards European integration.

9. 50% of the voters are having good political knowledge which means the votes are given by the voters with knowledge on the parties and their positions.

Checking for null values in the data:

Count of null values for each row:

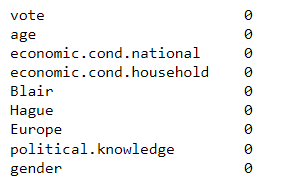


Figure 2: Counts of null values for columns in the dataset

* It indicates that there are no null values in the dataset.

Checking for duplicated rows in the dataset:

* There are 8 duplicate rows in the dataset. They are removed for further analysis.
* The resultant dataset now has 1517 rows and 9 columns.

Skewness in the data:

* The only continuous variable ‘Age’ in the dataset has been checked for outliers and skewness.

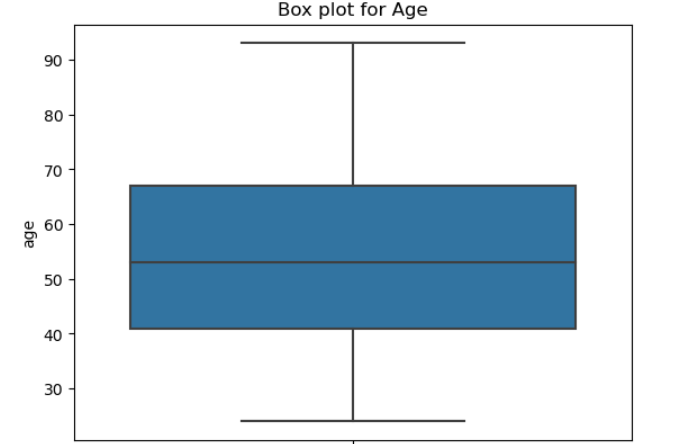


Figure 3: Box plot for Age

* The data for this feature does not have any outliers and is also not noticeably skewed towards any end. The age of the voters ranges between 21 and 93 where majority of the voters are within 41 to 67 years of Age.

**1.2 Perform EDA. Univariate, Bivariate and Multivariate analysis. Also check for outliers. Interpret the inferences for each.**

Ans:

Null value check:

* As discussed in the previous answer, there are no null values in any of the columns in the dataset. Hence, there is no need for any imputation.

Pre processing:

* The columns 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge' are ordinal variables as they represent scores according to business understanding.
* The data types for the columns are integers but since they are categorical in nature, converting the data types of the column to categorical.

Updated basic information of the dataset:

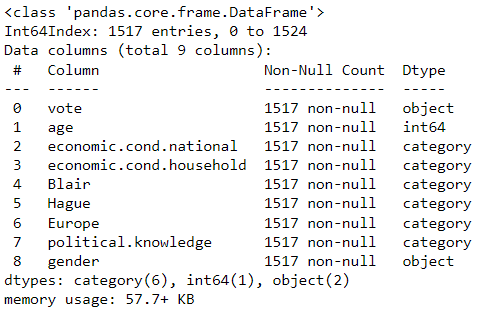
****

Figure 4: Information of the dataset after data pre-processing

Shape of the dataset:

* There are 1517 rows and 9 columns in the dataset after removing duplicates and performing preprocessing steps.

Univariate analysis:

Box plot for Age:

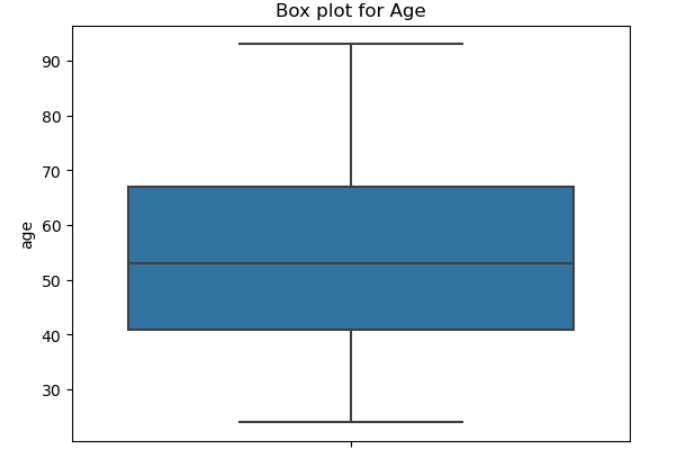


Figure 5: Box plot for Age

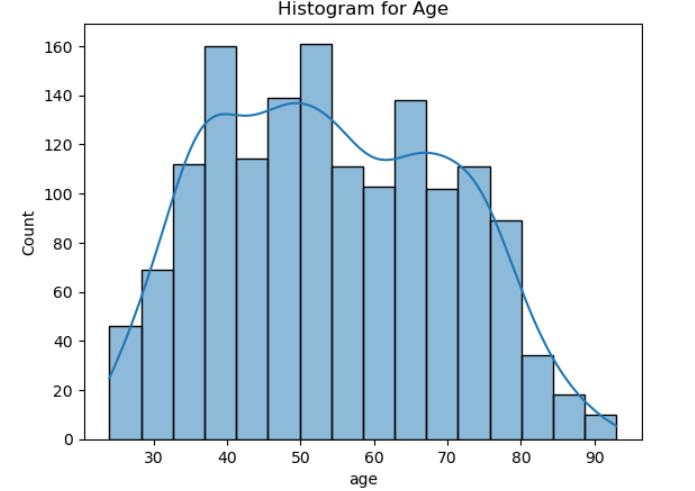
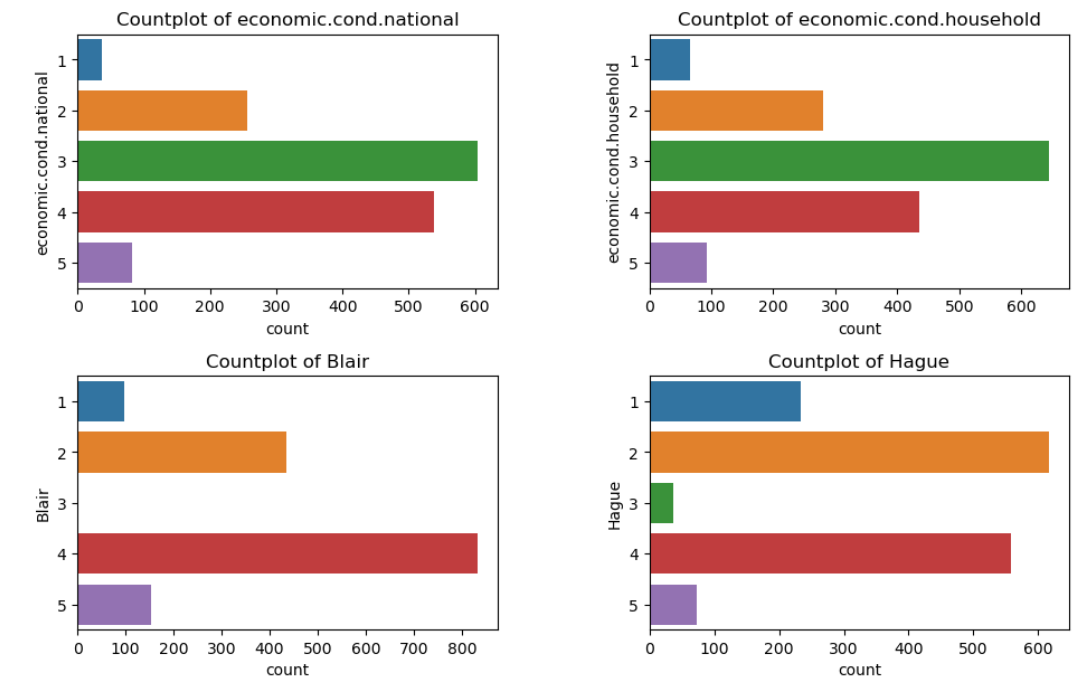


Figure 6: Histogram for Age

Insights:

1. There are no outliers in ‘Age’ column and the data is not very skewed to any end.
2. Most of the people who participated in voting are within 41 - 67 years.
3. The distribution of Age is not perfectly normal and has multiple peaks.

Count plots of categorical variables:



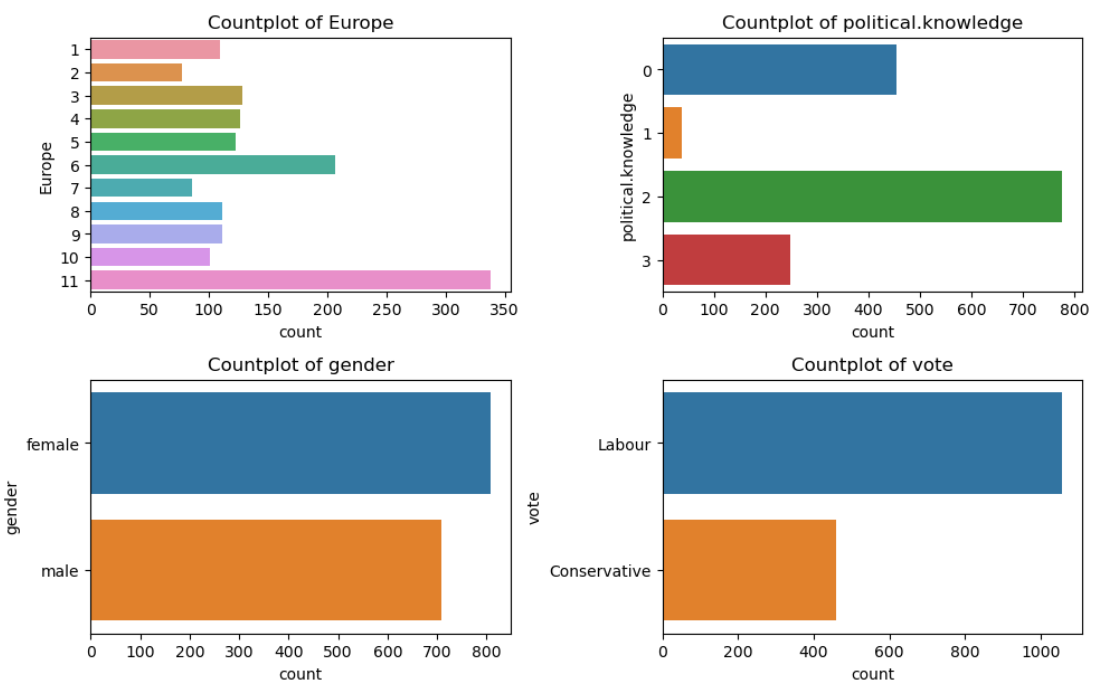


Figure 7: Count plots for categorical variables

Value counts for each categorical feature:

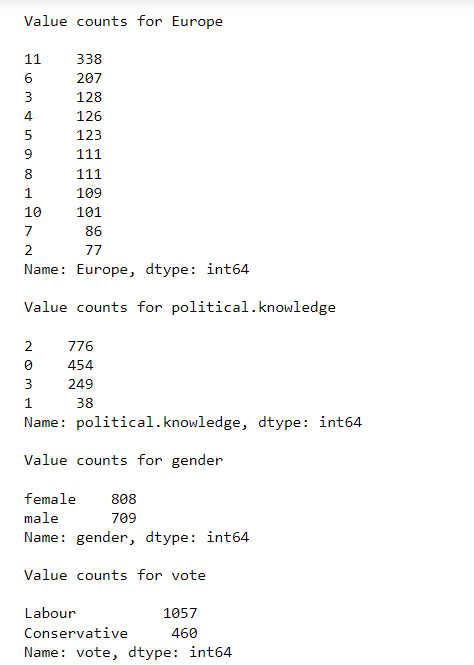
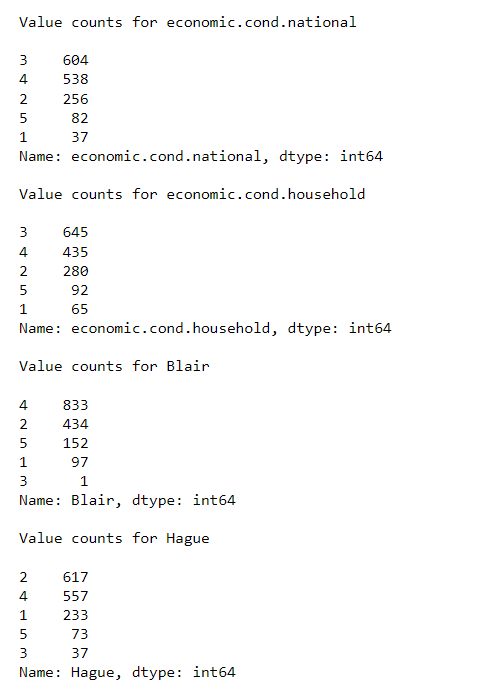
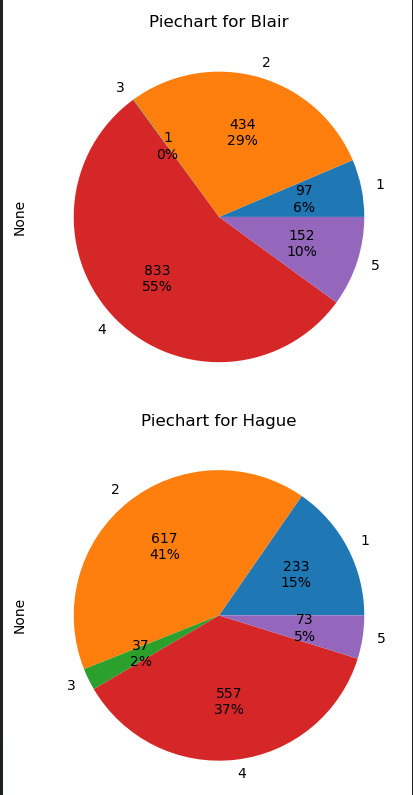
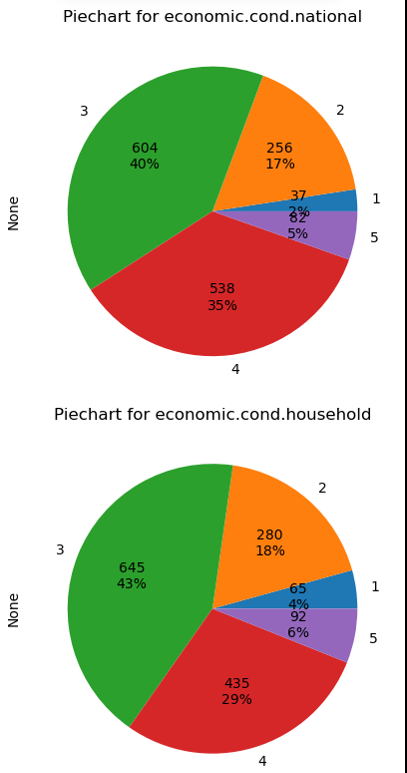


Figure 8: Value counts for all categorical variables

Pie Charts for each categorical variable:



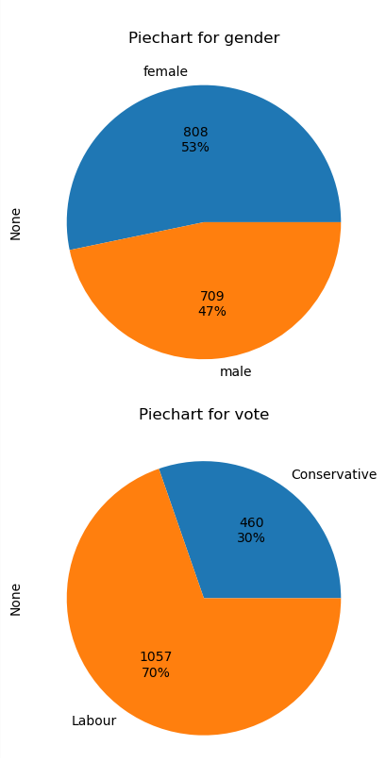
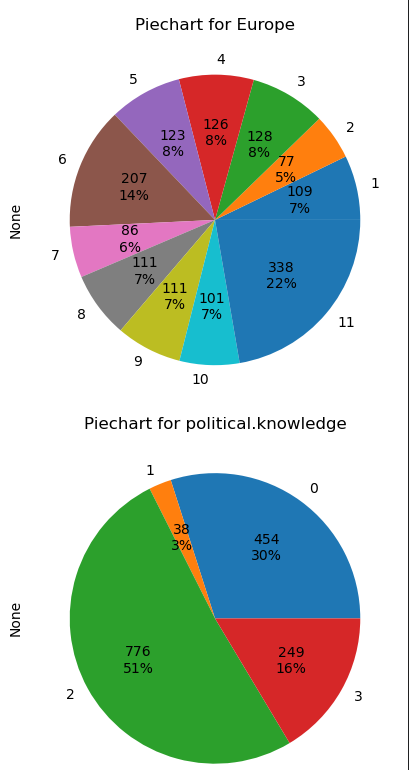


Figure 9: Pie charts for all categorical variables

Insights:

1. Around 75%(highest ratio) of the people have assessed current national and household economic conditions to be 3 and 4.
2. At Least 65% of the voters have given an assessment score of greater than 4 to Blair.
3. Around 58% of the voters have given an assessment score of less than 3 to Hague.
4. 22% of the people have the highest eurosceptic sentiment.
5. 30% of the voters seem to have no political knowledge whereas 67% have political knowledge of 2 or more on a scale of 0 to 3.
6. 53% of voters are female and 47% of the voters are male.
7. 70% of the voters have voted for Labour party and 30% voted for Conservative party.

Bivariate analysis:

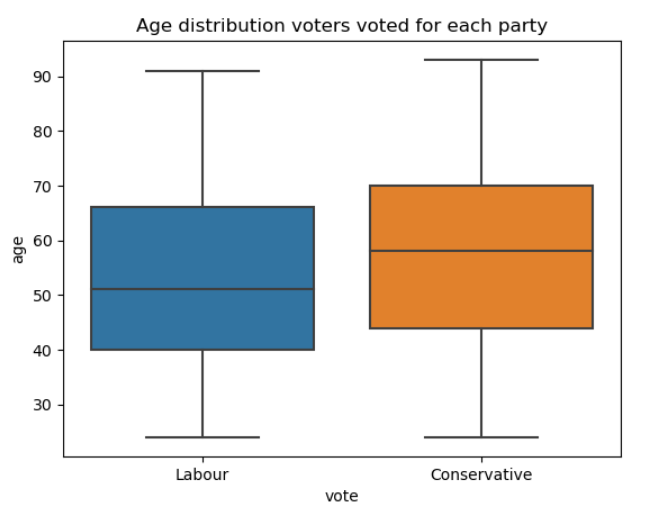
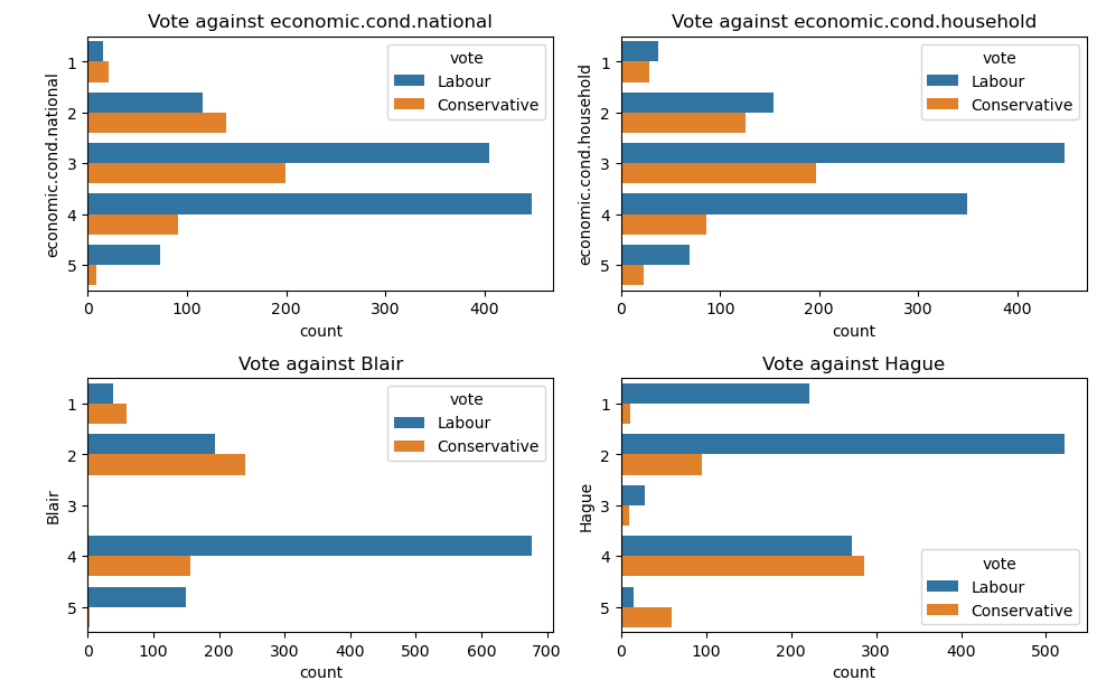


Figure 10: Age distribution of voters voted for each party

Insights:

1. The median age of voters voted for Conservative party is slightly higher than the median age of voters who voted for Labour party.

Count plots of various features of voters who voted for each party:



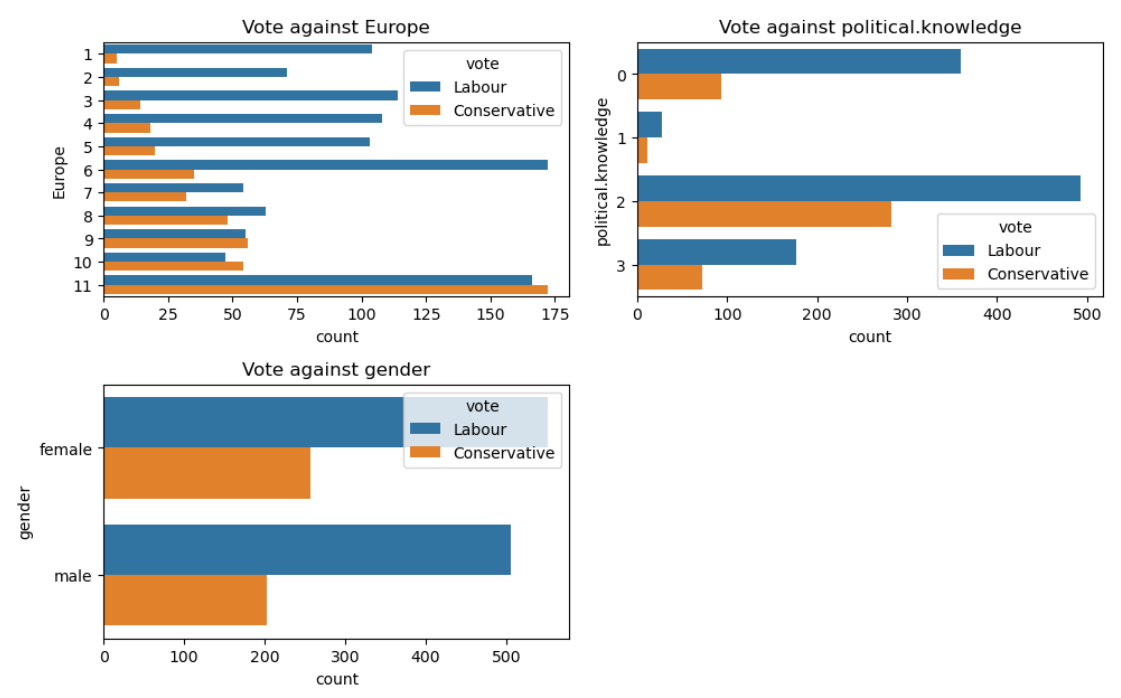


Figure 11: Count plots of all categorical variables distinguished by party voted for

Insights:

1. Majority of the voters who voted for Labour party have assessed current national economic conditions to be between 3 and 4
2. Whereas the voters who voted for conservative party have assessed current national and household economic conditions to be between 2 and 3.
3. It is evident from the above graphs that voters who have assessed a particular party leader with higher score have voted for the same party.
4. Most of the people who voted for Labour party have political knowledge of 0 and 2. Whereas, a considerable percentage of voters who have political knowledge of greater than 2 have also voted for Conservative party.

Multivariate analysis:

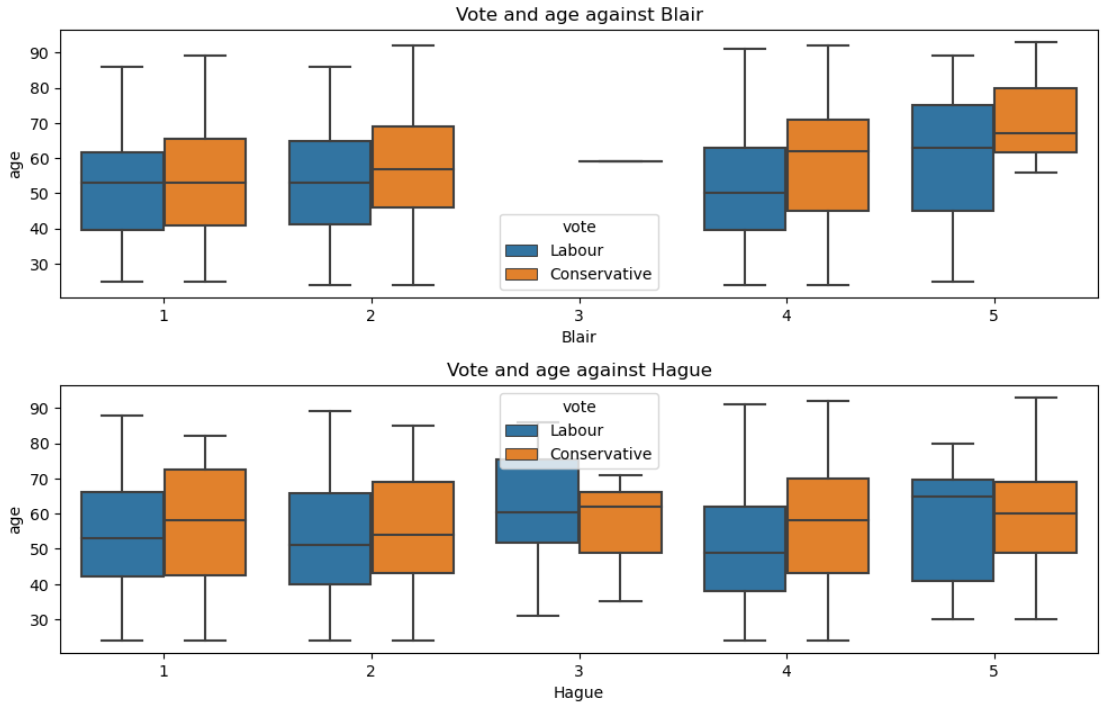
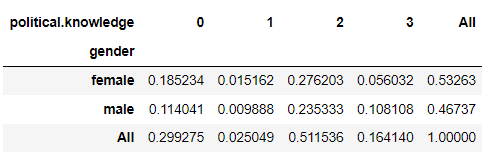
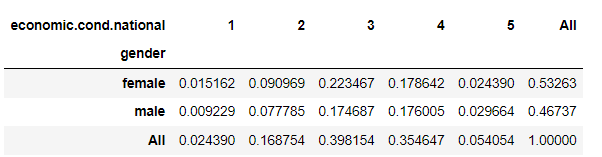
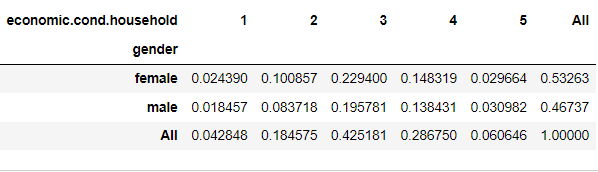


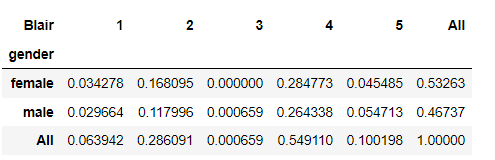
Figure 12: Age distribution of people with respect to assessment scores given to leaders

Percentages of voters voted to different parties w.r.t Gender and all other categorical variables









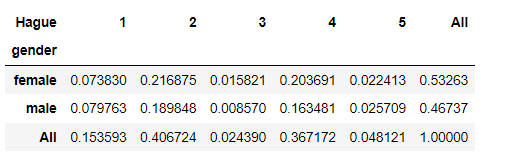


Figure 13: Percentages of voters voted to different parties w.r.t Gender and all other categorical variables

Insights:

1. Around 2.2% and 2.5% of females and males respectively have given a score of 5 for Hague.
2. Around 7.3% and 7.9% of females and males have given a score of 1 for Hague respectively.
3. Around 28% and 26% of females and males respectively have given a score of 4 to Blair.

**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). Data split, ratio defined for the split, train-test split should be discussed**

Ans:

Encoding:

* Age being a continuous column need not be encoded.
* The columns 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge' are ordinal variables. They are already in numerical encoded format hence no encoding is needed for them.
* Categorical variable ‘gender’ has been encoded using one-hot encoding.
* To interpret probabilities better we have replaced the ‘Labour’ column with 1 and ‘Conservative’ column with 0. We will proceed with further modeling by applying encoding to ‘gender’ column

Basic information of the dataset after encoding:

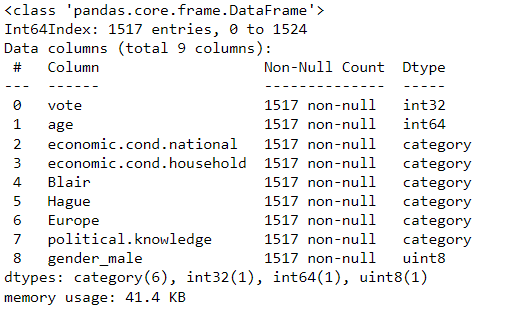


Figure 14: Basic information of encoded dataset

Viewing the first 5 rows of encoded dataset:

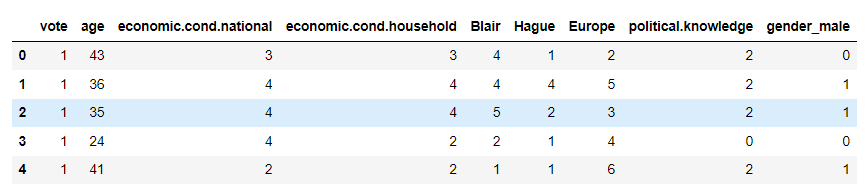


Table 5: Top 5 rows of encoded dataset

Scaling:

* As seen from the table above, most of the categorical variables are in the same scale.
* Age, which is a continuous variable, is also on a comparable scale with the magnitudes of all other categorical variables.
* The mean and standard deviation of all other categorical variables is around single digits i.e., ones place
* The mean and standard deviation of Age is around double digits and it mostly does not cross 100 i.e., tens place.
* These indicate that the magnitudes of all the independent features are comparable with each other and no scaling is required. The data should be able to work while using distance based algorithms as well.

Data split:

* The target variable has been dropped from the original dataset.
* Another dataframe with only the target variable has been considered.
* The data has been split into training and testing data with a 70: 30 ratio.
* Stratify parameter has been used so that the categories of the target class get distributed into training and testing data without any bias.
* Random state is considered to be 1.

Viewing the first rows of training dataset:

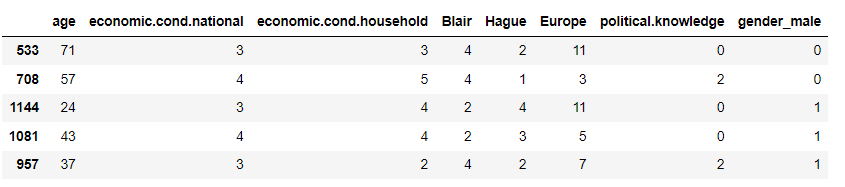


Table 6: Top 5 rows of independent variables of training dataset

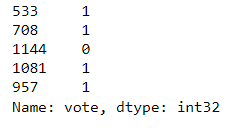


Table 7: Top 5 rows of dependent variables of training dataset

Viewing the first rows of the testing dataset:

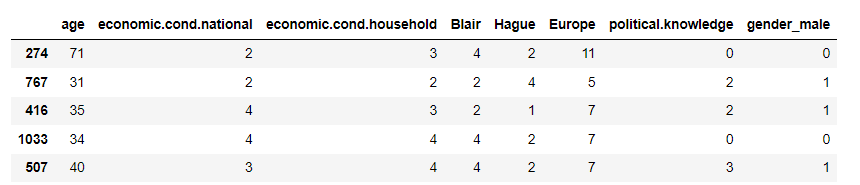


Table 8: Top 5 rows of independent variables of testing dataset

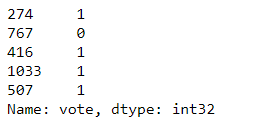


Table 9: Top 5 rows of dependent variables of testing dataset

**1.4 Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models. Successful implementation of each model. Logical reasons should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

**Ans:**

Logistic regression:

Building model:

1. Split the data into training and test data in 70:30 ratio
2. Target variable separated out is ‘vote’.
3. Building a Logistic Regression model with no parameters defined explicitly on the training data.
4. Default parameters of the model are:
   1. Penalty: ‘l2’ (add a l2 penalty term)
   2. Solver: ‘lfgbs’
5. The built model has been passed for prediction.
6. A glimpse of the predicted probabilities for training and test data:

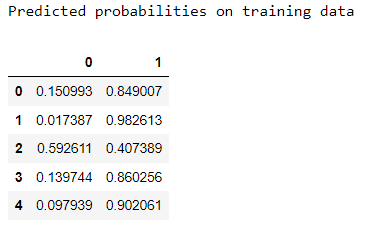
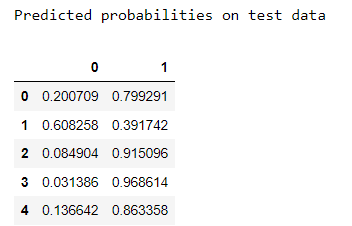
 

Figure 15: Predicted probabilities on training and test data after applying Logistic Regression

Accuracy of the model on training data: 82.85%

Accuracy of the model on testing data: 85.53%

* The model has performed well on training and testing data well. In Fact the accuracy on the test data is more than the accuracy of the training data. There is not much difference in the accuracies of training and test data. It means the model has not over fit or under fit and has generalized well.

Linear Discriminant Analysis:

Steps followed to perform Linear Discriminant Analysis:

1. Split the data into training and test data in 70:30 ratio
2. Target variable separated out is ‘vote’.
3. Applying LinearDiscriminantAnalysis from scikit learn on the training data set.
4. The obtained coefficients after performing LDA are:

[-0.02, 0.36, 0.03, 0.69, -0.97, -0.22, -0.48, 0.02]

1. The intercept obtained from the model on training data: 3.68
2. The linear discriminant function built by LDA model on training data set is as follows:

3.68 + (-0.02) \* age + (0.36) \* economic.cond.national + (0.03) \* economic,cond.household + (0.69) \* Blair + (-0.97) \* Hague + (-0.22) \* Europe + (-0.48) \* political\_knowledge + (0.02) \* gender\_male

1. Inferences from above LDA equation:
   1. High value of national economic condition, high value of Eurosceptic sentiment by the voter and high assessment given to Blare by the voter result the voter voting to the ‘Labour’ party.
   2. On the other hand, a high score to Hague results in a high probability that the voter votes to Conservative party.

Accuracy of the model on training data: 82.28%

Accuracy of the model on testing data: 85.31%

* LDA model has performed comparatively equal to the logistic regression model. There is not much difference in the accuracies of training and test data. It means the model has not over fit or under fit and has generalized well.

**1.5 Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model. Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

Gaussian NaiveBayes model:

1. Using NaiveBayes classifier on the training data set with default parameters.
2. The default parameter used for Gaussian Naive Bayes is var\_smoothing = 0.000000009
3. This value essentially adds weights to user defined values which smoothens the Gaussian curve and lets the values far away from the curve also be given equal weight.
4. Accuracy of the model on training dataset: 82%
5. Accuracy of the model on the testing dataset: 85.74%
6. The model has performed well on training dataset and testing dataset and as we can see the performance on training dataset is less than that in the testing dataset.
7. This can be because the class labels of the target variable are only 2 and are imbalance.

Applying SMOTE and trying various hyper parameters to improve accuracy of Gaussian Naive Bayes algorithm:

1. SMOTE has been applied on the training dataset to make the classes balance.
2. Various values for the hyper parameters were considered. If var\_smoothing is increased, the accuracy of the model decreases. Eg: If var\_smoothing is set to 1, accuracy decreases to around 70%
3. Considering values closer to default value for var\_smoothing, we get higher accuracy.
4. Applying Gaussian Naive Bayes on SMOTE trained dataset with var\_smoothing set to 0000000001.
5. Accuracy of the training dataset turned out to be 81.05%
6. Accuracy of the testing dataset has turned out to be 82.23%

Inferences from models:

1. We observe that the difference between training and test dataset upon applying SMOTE has decreased.
2. This indicates that the model has not either overfit or underfit. The model has generalize well.
3. The model also gives high accuracies of 81% - 82%

KNN model:

1. Using K Nearest Neighbors on the training data with default parameters.
2. The default parameters are n\_neighbors**=5,** weights**='uniform',** algorithm**='auto',** leaf\_size**=30,** p**=2,** metric**='minkowski',**
3. In this model, we consider values of 5 nearest neighbors to predict the outcome of the test data. Minkowski distance is used as a distance calculation metric. But the p-value is 2 which means Euclidean distance is used as a default distance calculation metric.
4. Accuracy of the model on training dataset: 84.91%
5. Accuracy of the model on test dataset: 80.7%
6. The model has predicted the data with higher accuracy compared to Naive Bayes algorithm. Training accuracy and test accuracy are considerably near to each other which means that the model works well on training and test data. The model has generalized well.

Applying KNN with different set of hyperparameters:

1. We will increase the nearest neighbors and also change the distance calculation metric
2. Applying K Nearest neighbors with the following hyper parameters: n\_neighbors = 7, leaf\_size = 30, p = 1
3. We have increased the leaf size, maximum number of neighbors to be considered, and Minkowski distance.
4. Accuracy on the training data with the above set of hyper parameters is 84%
5. Accuracy on the test data with the above set of hyper parameters in 83%

Inferences:

1. By increasing the number of neighbors, the accuracy on the test data for this hyper parameter tuned algorithm is higher than that of the original.
2. Accuracy of the training and test datasets have come closer to each other which means the model is improving with hyper parameter tuning.
3. Hence, KNN with the above parameters gives better accuracy on the given dataset.

**1.6 Model Tuning, Bagging and Boosting . Apply grid search on each model (include all models) and make models on best\_params. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.**

Using SMOTE on the training dataset:

* From the EDA of this dataset, we learned that there are around 1000 features with target variable as ‘Labour’ and around 400 variables with target variable as ‘Conservative’.
* This indicates that there is class imbalance and we can treat this class imbalance on the training dataset by using SMOTE.
* SMOTE is an oversampling technique that creates dummy variables based on the existing data.
* Shape of original dataset (1061, 8)
* Shape of SMOTE sampled dataset (1478, 8)
* Value counts of target variable before applying SMOTE
  + 1 : 739 and 0 : 322
* Value counts of target variable after applying SMOTE
  + 0 : 739 and 1 : 739

Hyper parameter tuning on Logistic Regression model with SMOTE sampled data:

1. Parameters considered:
   1. Penalty: ‘l2’, ‘none’
   2. Solver: lfgbs, ‘newton-cg’
2. Applied grid search CV on logistic regression model with following sets of hyper parameters:
   1. GridSearchCV(cv=3, estimator=LogisticRegression(n\_jobs=-1), n\_jobs=-1,
   2. param\_grid={'penalty': ['l2', 'none'],
   3. 'solver': ['lfgbs', 'newton-cg']},
   4. scoring='f1')
3. The best parameters found for Logistic Regression by using GridSearch CV are: LogisticRegression(n\_jobs=-1, solver='newton-cg'). Used 3 fold cross validation to find the best parameter.
4. Accuracy of training dataset: 81.79%
5. Accuracy of the testing dataset: 82.89%
6. Accuracy of the training and test datasets have come closer to each other which means the model is improving with the best estimators obtained from hyper parameter tuning.

Hyper parameter tuning on Gaussian Naive Bayes model with SMOTE sampled data:

1. Parameters considered: Var\_smoothing: 0.000001, 0.00001, 0.0001, 0.001, 0.1, 0
2. 3 fold cross validation used by GridSearchCV.
3. Best parameters found out by grid search CV: var\_smoothing = 0.000001
4. Applying Gaussian NaiveBayes using the best estimator built by GridSearchCV
5. Accuracy on the training dataset: 81.05%
6. Accuracy on the testing dataset: 82.24%
7. Accuracy of the training and test datasets have come closer to each other which means the model is improving with the best estimators obtained from hyper parameter tuning.

Hyper parameter tuning on KNN model with SMOTE sampled data:

1. Parameters considered: params = {'n\_neighbors':[3,4,5,6,7], 'leaf\_size':20,21,...,40, 'weights':['uniform', 'distance'], 'p': [0,1,2]}
2. Applying GridSearchCV, best estimators obtained are: leaf\_size=20, n\_neighbors=3, weights= ‘distance’
3. Applying KNN on the data with the best parameters.
4. Accuracy of the model on training data: 99.9%
5. Accuracy of the model on testing data: 78.51%
6. Training data accuracy is very high whereas the test data accuracy is low and there is significant difference between both accuracies. Model has overfitted.

Hyper parameter tuning on Bagging Classifier with SMOTE sampled data:

1. Bagging Classifier with Base estimator used: CART
2. Parameters: 'n\_estimators' : [50, 60, 70, 80, 90, 100], 'max\_samples': [1, 5, 10, 15, 20], 'max\_features': [1, 10, 20, 30]
3. Applying grid search CV on the model with the above base estimator and given parameters.
4. Best parameters obtained from GridSearchCV are: max\_features=1,max\_samples=15, n\_estimators=80)
5. The model is constructed using the best parameters obtained and the training dataset has been fit.
6. Accuracy of the model for training data: 77.06
7. Accuracy of the model on testing data: 82.67
8. Training accuracy has dropped by using Bagging Classifier with a random state of 1. Hence we will choose other models which show better accuracy.

Hyper parameter tuning on Random Forest Classifier with SMOTE sampled data::

1. GridSearchCV is applied on RandomForestClassifier with parameters:
   1. {'n\_estimators' : [50, 60, 70, 80, 90, 100],
   2. 'min\_samples\_split' : [10, 20, 30],
   3. 'min\_samples\_leaf' : [10, 20, 30]}
2. The best estimator function turned out to be min\_samples\_leaf=10, min\_samples\_split=30, n\_estimators=50, random\_state = 1
3. Applying the RandomForestClassifier with above parameters.
4. Top 5 important independent variables according to feature importances are

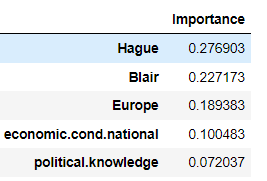


Figure 16: Feature importances from Random Forest Classifier

1. Accuracy of the model on training data: 85.79
2. Accuracy of the model on testing data: 84.87
3. This model seems to show the best training and test accuracy compared to all other models. The most important features according to this model are Hague, Blair, Europe.

Hyper parameter tuning on AdaBoost Classifier with SMOTE sampled data:

1. GridSearchCV is applied on Adaptive boosting with parameters:
   1. 'n\_estimators' : [50, 60, 70, 80, 90, 100]
2. The best estimator function turned out to be n\_estimators=90
3. Applying the Adaptive boosting with above parameters.
4. Feature importances according to Adaboost classifier

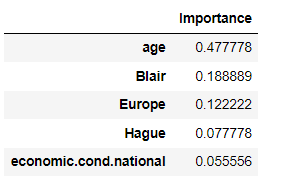


Figure 17: Feature importances from Adaptive Boosting Classifier

1. Accuracy of the model on training data: 84.2
2. Accuracy of the model on testing data: 83.3
3. Adaptive boosting algorithm has also provided better training and test accuracies, Most important features according to this model are Age, Blair, Europe and Hague.

Hyper parameter tuning on GradientBoost Classifier with SMOTE sampled data::

1. GridSearchCV is applied on Gradient boosting with parameters: 'n\_estimators' : [50, 60, 70, 80, 90, 100]
2. The best estimator function turned out to be n\_estimators=70
3. Applying the Adaptive boosting with above parameters.
4. Feature importances according to Gradient boosting model:

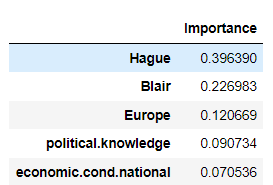


Figure 18: Feature importances from Gradient Boost classifier

1. Accuracy of the model on training data: 87.7%
2. Accuracy of the model on testing data: 82.67%
3. Gradient Boosting has given less accuracies on test data and high on training data compared to other models, this model seems to be slightly overfit.

**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, classification report. Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.**

**Ans:**

Evaluating performance metrics for different models built:

Description of different parameters considered in evaluating the performance of a machine learning model:

To actually classify the predictor variable, the algorithm calculates the probabilities of the predictor variable belonging to each class. A threshold needs to be devised to partition the response space into success and failure. Typically, the threshold is set at 50% level.

If the probability of success is 50% or above for a given combination of predictors, the value of response is taken to be 1, otherwise 0.

However, this threshold may be set at some other convenient level.

Classification matrix:

Let P be the total number of successes (positives) in the data and N be the total number of failures (negatives).

If a success is predicted as success, it is an example of True Positive (TP).

If on the other hand a failure is predicted as failure, it is an example of True Negative (TN).

In both cases, classification is correct.

However,

if a success is predicted as a failure, it is an example of False Negative (FN)

if a failure is predicted as a success, it is an example of false positives. These are misclassified.

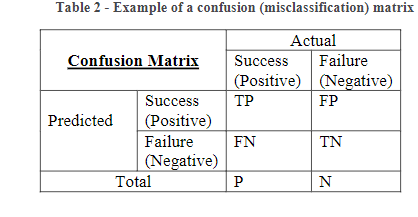


Table 10: Confusion matrix

Probability of misclassification = 𝐹𝑁+𝐹𝑃 𝑛 ,

where n is the sample size.

For the perfect model, misclassification probability is 0; i.e. no observation would have been misclassified.

This indicates overfit of the model and not to be recommended, since such a model will not have good predicting power.

A few other performance metrics are equally important.

Precision = 𝑇𝑃 /𝑇𝑃+𝐹𝑃, i.e. among all the successes (positives) in the data, how many are identified as positive by the model.

Specificity = 𝑇𝑁 /𝑇𝑁+𝐹𝑃, i.e. among all failures (negatives) in the data, how many are actually identified as negative by the model.

Sensitivity or Recall = 𝑇𝑃/ 𝑇𝑃+𝐹𝑁 , i.e. among all the predicted successes, how many are actually success.

The F-score of the model is defined as 2(𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛∗𝑅𝑒𝑐𝑎𝑙𝑙) / 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛+𝑅𝑒𝑐𝑎𝑙𝑙 .

F is between 0 and 1, and the closer it is to 1, the better is the model.

1. Performance metrics for tuned Logistic regression model:
2. Predicted probabilities:
   1. On training data: On test data:

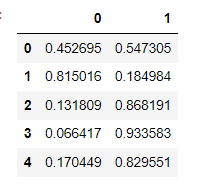
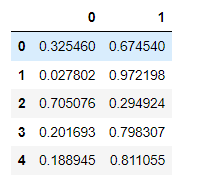


Figure 19: Predicted probabilities – Logistic Regression

* 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.45 whereas the probability of predictor variable belonging to class 1 is 0.54

1. Accuracy:
   1. Accuracy of the logistic regression model on the training data: 81.8%
   2. Accuracy of the logistic regression model on the test data: 82.9%
   3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
2. Confusion matrix:
   1. Confusion matrix on training and test data:

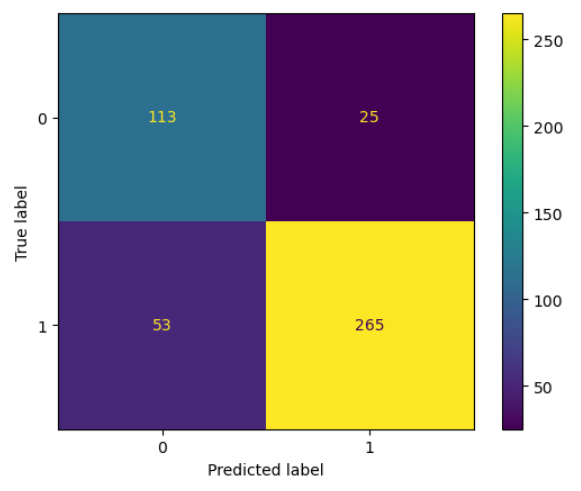
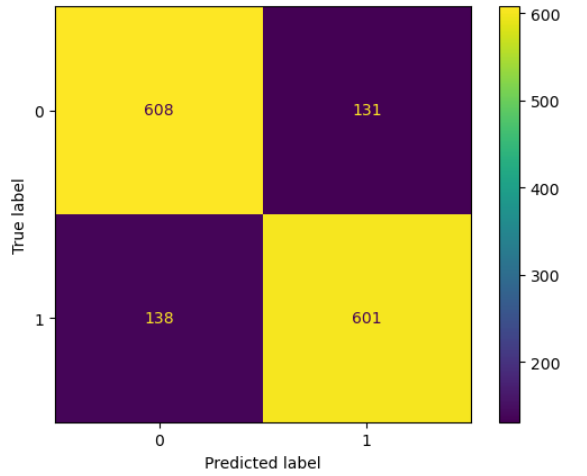


Figure 20: Confusion matrix – train and test data (Logistic Regression)

* 1. Insights on test data:
     1. 113 people have correctly been predicted to vote of Conservative party. 265 have correctly been predicted to vote for Labour party.
     2. 25 people have been incorrectly predicted to vote for Labour party.
     3. 53 people have been incorrectly predicted to vote for Conservative party.

1. Classification report for training data and testing data:

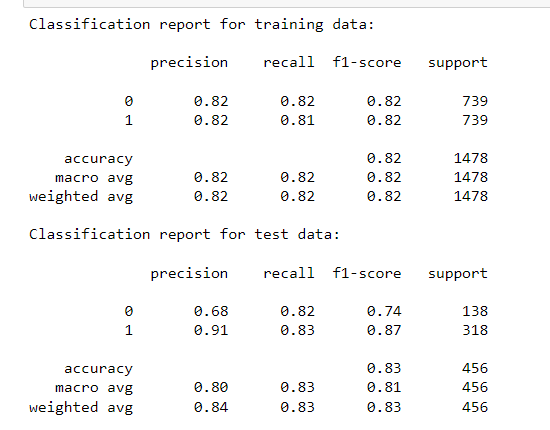


Figure 21: Classification report – training and test data (Logistic Regression)

* 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
  2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.

1. ROC Curve on training data:
   1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis

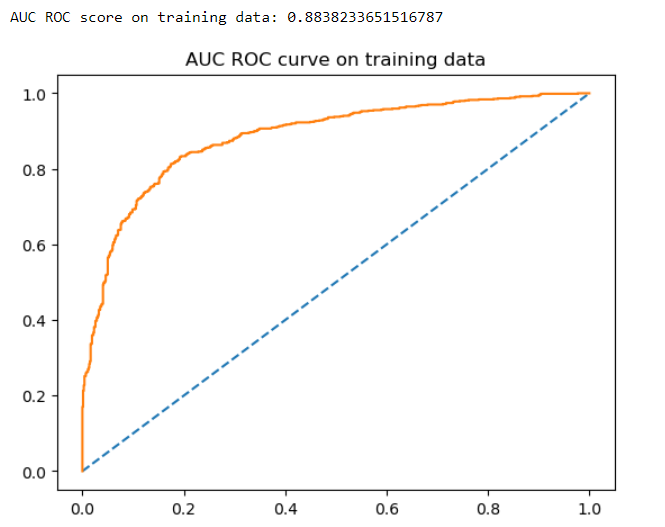


Figure 22: AUC ROC curve – training data (Logistic Regression)

* 1. Area under the ROC curve: AUC ROC score on training data: 0.88

1. ROC Curve for the test data:

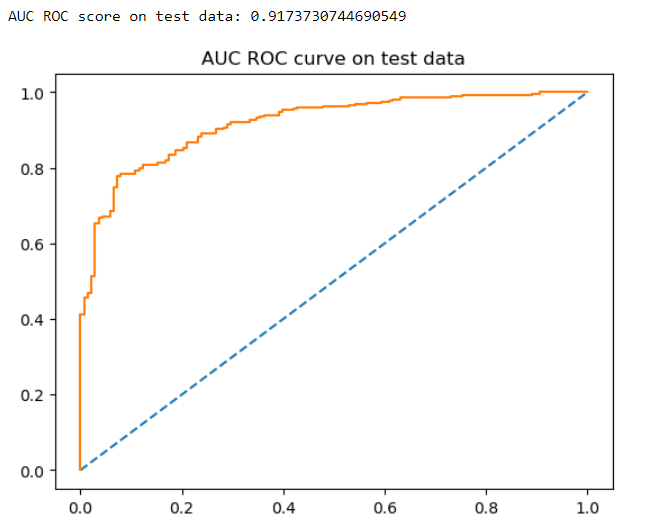


Figure 23: AUC ROC curve – testing data (Logistic Regression)

* 1. AUC ROC score on test data: 0.91
  2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned Naive Bayes model:
   1. Predicted probabilities:
      1. On training data: On test data:

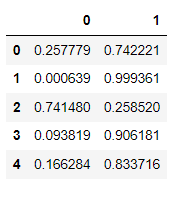
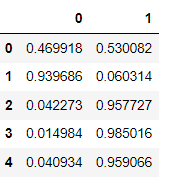
 

Figure 24: Predicted probabilities – Gaussian Naive Bayes

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.46 whereas the probability of predictor variable belonging to class 1 is 0.53
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 81%
     2. Accuracy of the logistic regression model on the test data: 82.23%
     3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
  2. Confusion matrix:
     1. Confusion matrix on training and test data:

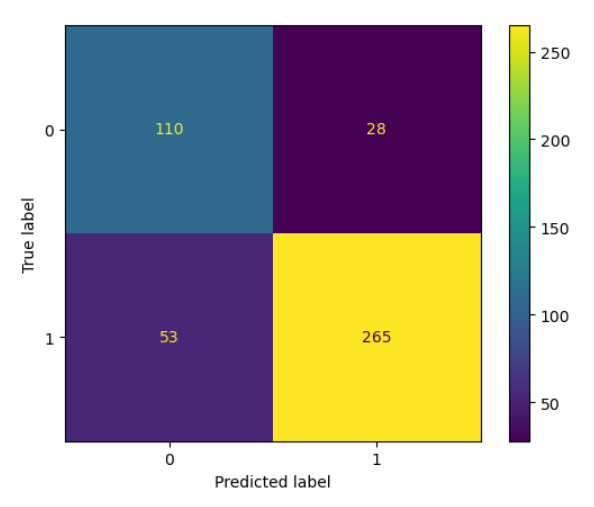
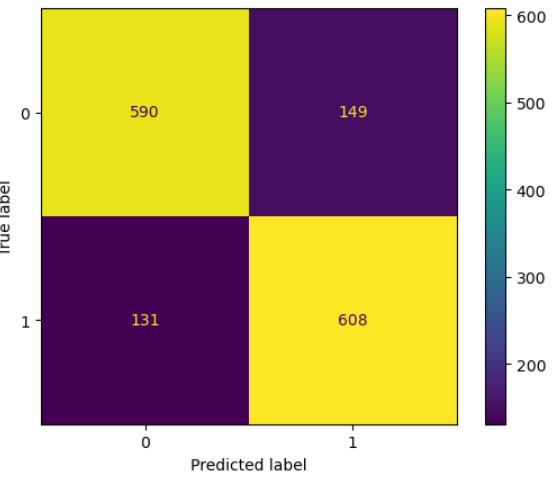


Figure 25: Confusion matrix – train and test data (Gaussian Naive Bayes)

* + 1. Insights on test data:
       1. 110 people have correctly been predicted to vote of Conservative party. 265 have correctly been predicted to vote for Labour party.
       2. 28 people have been incorrectly predicted to vote for Labour party.
       3. 53 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:

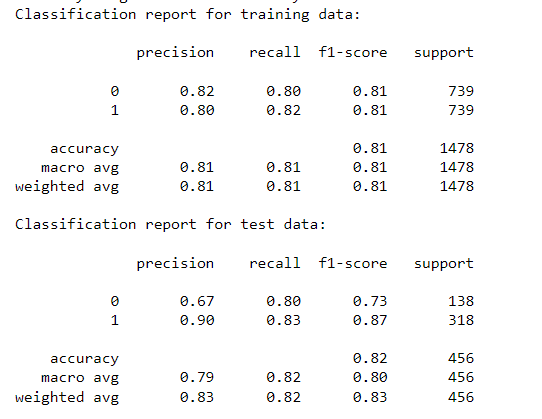


Figure 26: Classification report – training and test data (Gaussian Naive Bayes)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

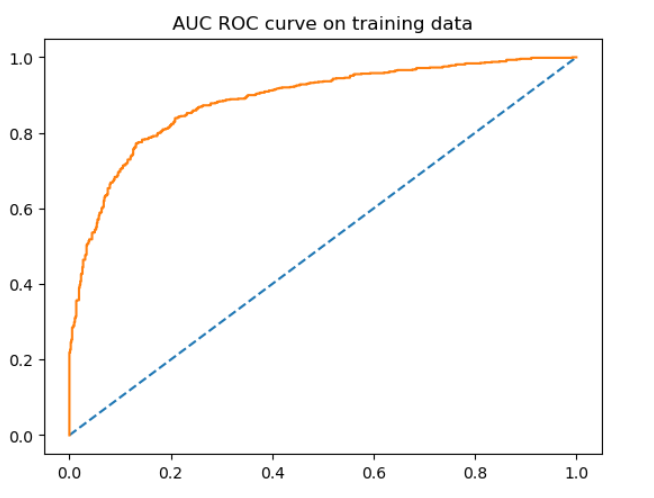


Figure 27: AUC ROC curve – training data (Gaussian Naive Bayes)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.88
  1. ROC Curve for the test data:

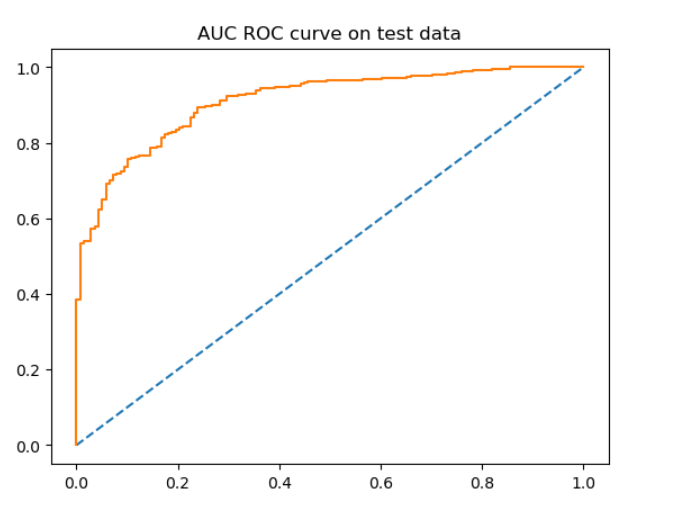


Figure 28: AUC ROC curve – testing data (Gaussian Naive Bayes)

* + 1. AUC ROC score on test data: 0.91
    2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned K Nearest Neighbors model:
   1. Predicted probabilities:
      1. On training data: On test data:

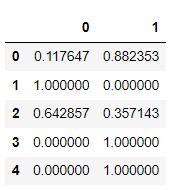
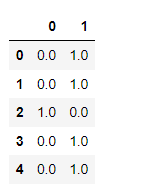


Figure 29: Predicted probabilities – K Nearest Neighbors

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.11 whereas the probability of predictor variable belonging to class 1 is 0.88
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 99%
     2. Accuracy of the logistic regression model on the test data: 78.5%
     3. Insights: Accuracies of training and test data are not close to each other which indicates that the model has overfitted.
  2. Confusion matrix:

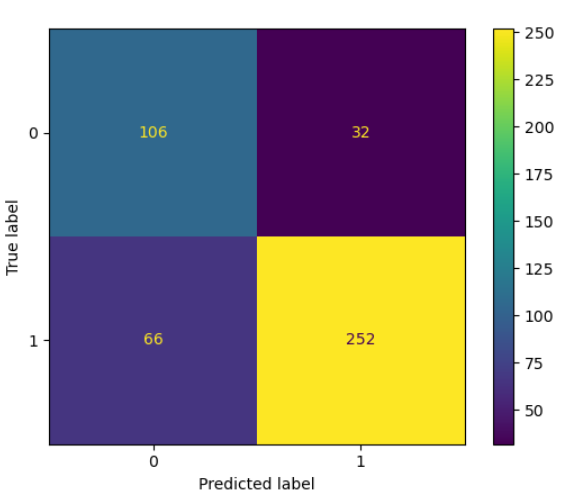
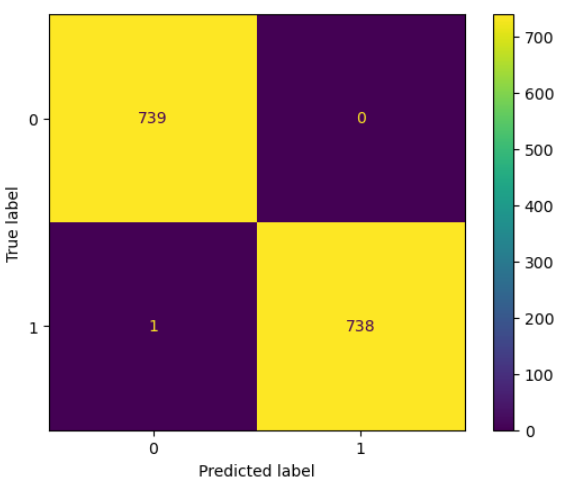


Figure 30: Confusion matrix – train and test data (K Nearest Neighbors)

* + 1. Insights on test data:
       1. 106 people have correctly been predicted to vote of Conservative party. 252 have correctly been predicted to vote for Labour party.
       2. 32 people have been incorrectly predicted to vote for Labour party.
       3. 66 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:
     1. Report on training data:

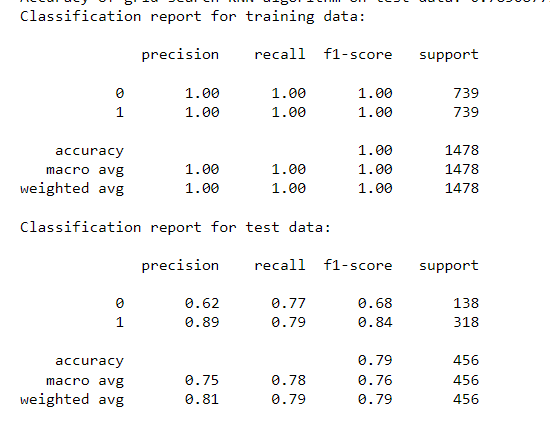


Figure 31: Classification report – training and test data (K Nearest Neighbors)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

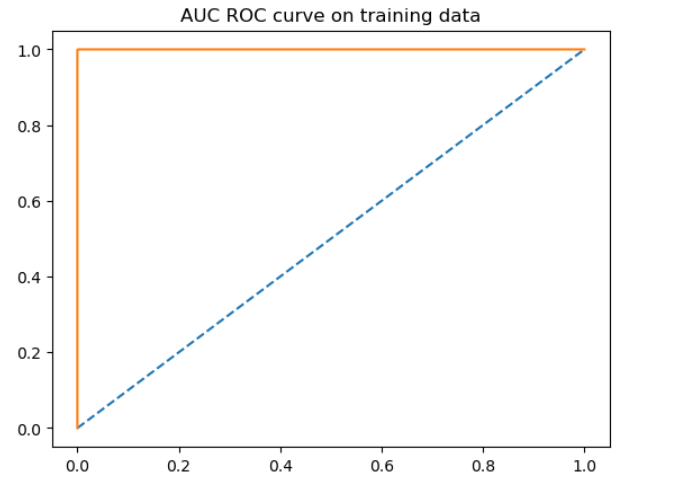


Figure 32: AUC ROC curve – training data (K Nearest Neighbors)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.99
  1. ROC Curve for the test data:

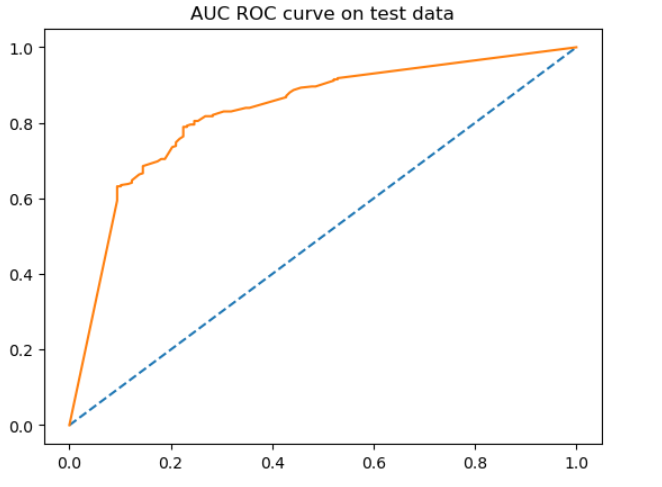


Figure 33: AUC ROC curve – testing data (K Nearest Neighbors)

* + 1. AUC ROC score on test data: 0.829
    2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned Bagging Classifier with base model as DecisionTreeClassifier model:
   1. Predicted probabilities:
      1. On training data: On test data:

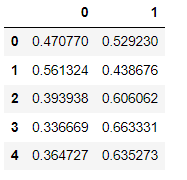
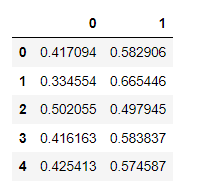


Figure 34: Predicted probabilities – Bagging Classifier

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.47 whereas the probability of predictor variable belonging to class 1 is 0.52
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 77%
     2. Accuracy of the logistic regression model on the test data: 82.67%
     3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
  2. Confusion matrix:

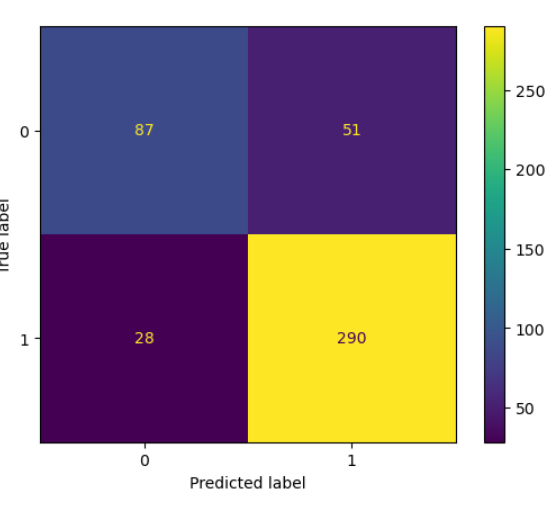
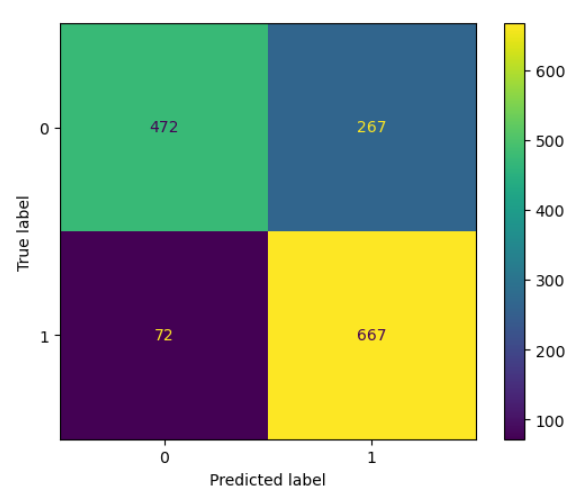


Figure 35: Confusion matrix – train and test data (Bagging Classifier)

* + 1. Insights on test data:
       1. 87 people have correctly been predicted to vote of Conservative party. 290 have correctly been predicted to vote for Labour party.
       2. 51 people have been incorrectly predicted to vote for Labour party.
       3. 28 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:
     1. Report on training data:

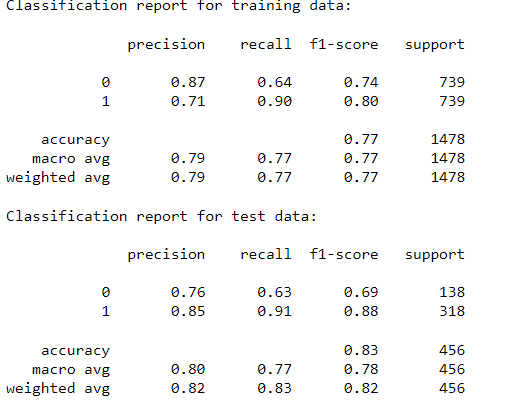


Figure 36: Classification report – training and test data (Bagging Classifier)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

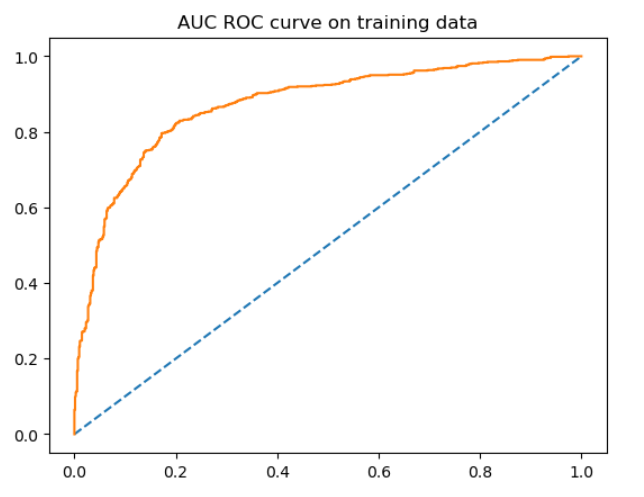


Figure 37: AUC ROC curve – training data (Bagging Classifier)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.87
  1. ROC Curve for the test data:

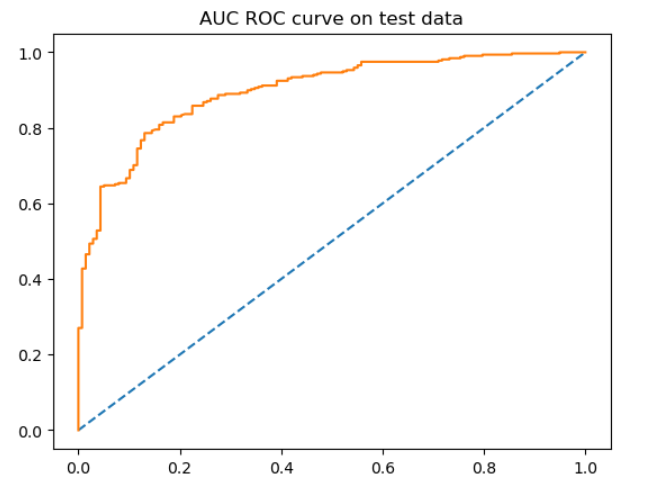


Figure 38: AUC ROC curve – testing data (Bagging Classifier)

* + 1. AUC ROC score on test data: 0.89
    2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned RandomForest Classifier model:
   1. Predicted probabilities:
      1. On training data: On test data:

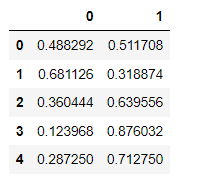
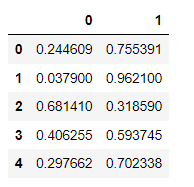


Figure 39: Predicted probabilities – Random Forest Classifier

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.48 whereas the probability of predictor variable belonging to class 1 is 0.51
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 85.79%
     2. Accuracy of the logistic regression model on the test data: 84.86%
     3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
  2. Confusion matrix:

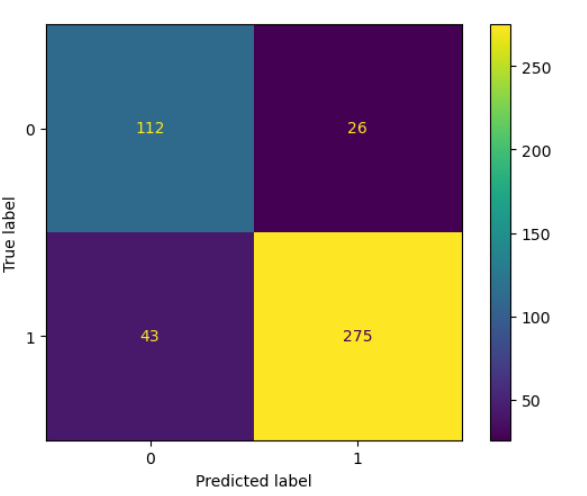
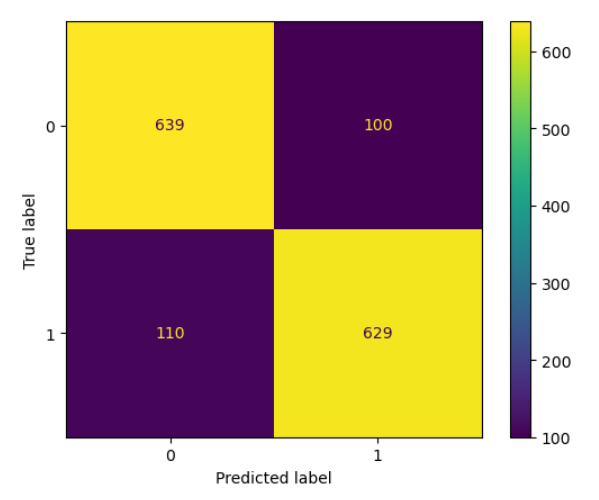


Figure 40: Confusion matrix – train and test data (Random Forest Classifier)

* + 1. Insights on test data:
       1. 112 people have correctly been predicted to vote of Conservative party. 275 have correctly been predicted to vote for Labour party.
       2. 26 people have been incorrectly predicted to vote for Labour party.
       3. 43 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:
     1. Report on training data:

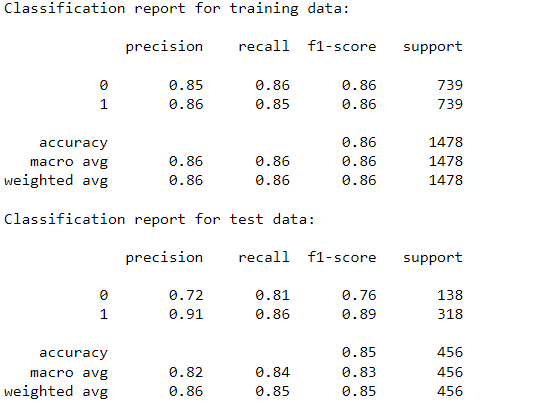


Figure 41: Classification report – training and test data (Random Forest Classifier)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

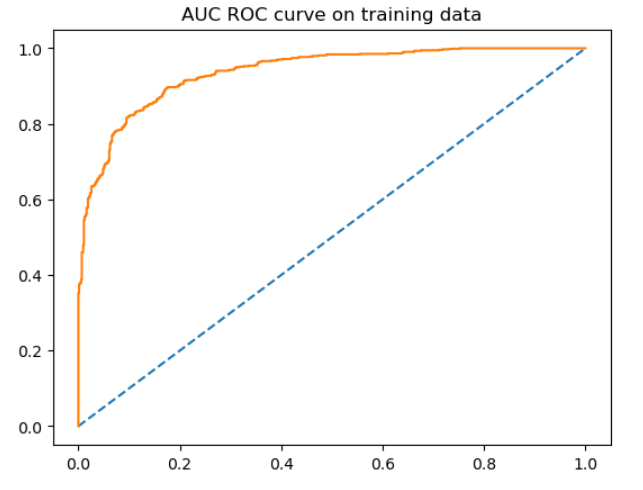


Figure 42: AUC ROC curve – training data (Random Forest Classifier)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.934
  1. ROC Curve for the test data:

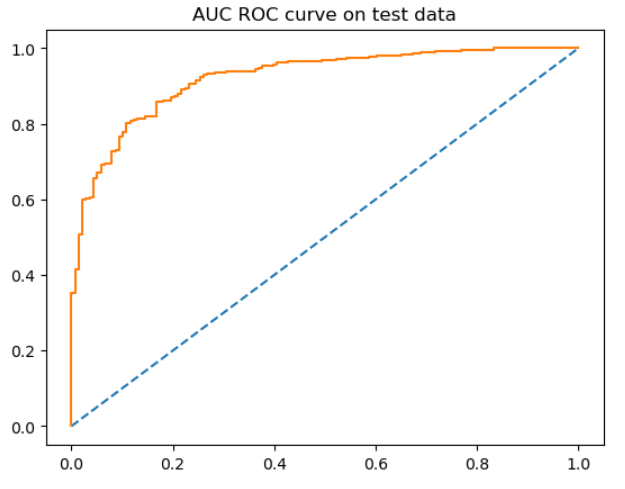


Figure 43: AUC ROC curve – testing data (Random Forest Classifier)

* + 1. AUC ROC score on test data: 0.919
    2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned Adaptive boosting model:
   1. Predicted probabilities:
      1. On training data: On test data:

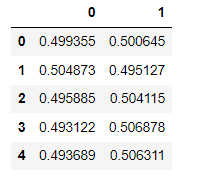
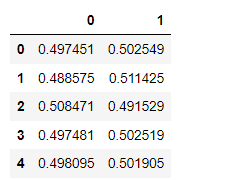


Figure 44: Predicted probabilities – Adaptive Boosting

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.49 whereas the probability of predictor variable belonging to class 1 is 0.50
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 84.23%
     2. Accuracy of the logistic regression model on the test data: 83.33%
     3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
  2. Confusion matrix:

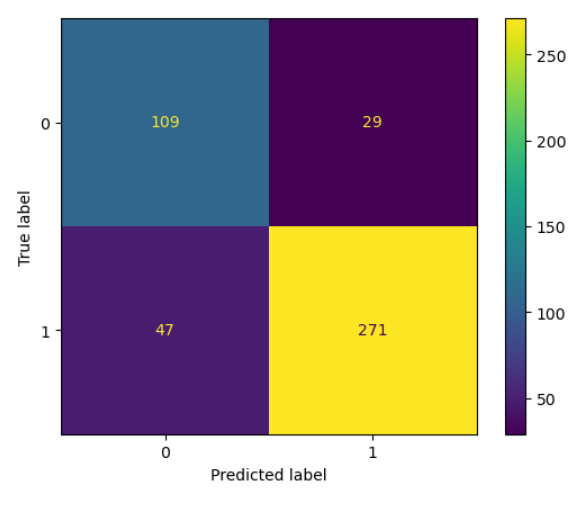
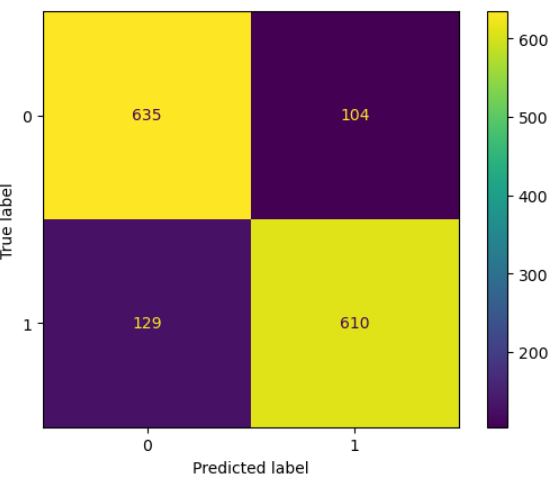


Figure 45: Confusion matrix – train and test data (Adaptive Boosting)

* + 1. Insights on test data:
       1. 109 people have correctly been predicted to vote of Conservative party. 271 have correctly been predicted to vote for Labour party.
       2. 29 people have been incorrectly predicted to vote for Labour party.
       3. 47 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:

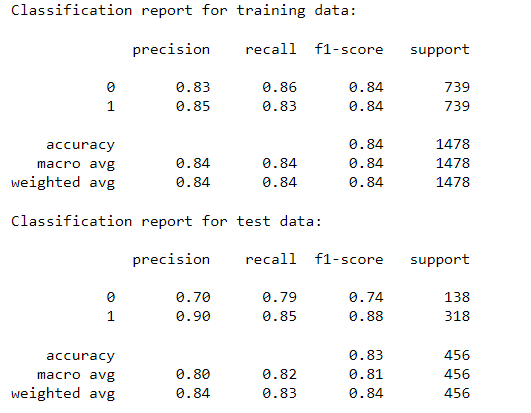


Figure 46: Classification report – training and test data (Adaptive Boosting)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

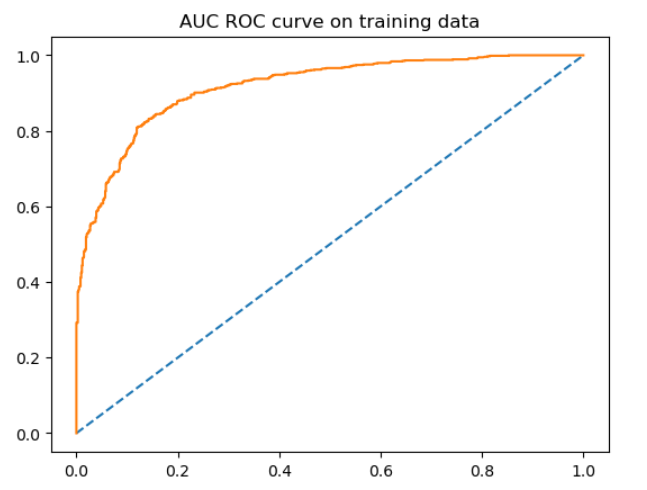


Figure 47: AUC ROC curve – training data (Adaptive Boosting)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.916
  1. ROC Curve for the test data:

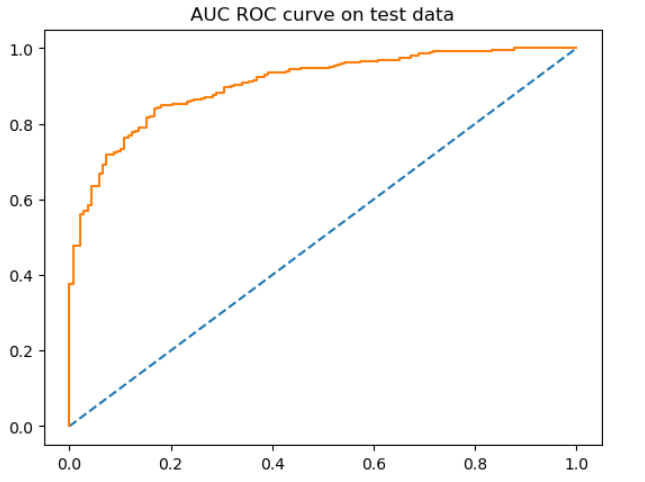


Figure 48: AUC ROC curve – testing data (Adaptive Boosting)

* + 1. AUC ROC score on test data: 0.905
    2. Higher the AUC ROC score tends to 1 better is the model.

1. Performance metrics for tuned Gradient boosting model:
   1. Predicted probabilities:
      1. On training data: On test data:

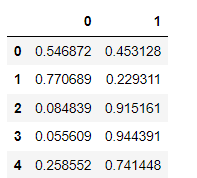
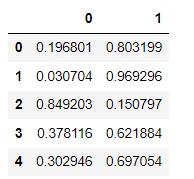


Figure 49: Predicted probabilities – Gradient Boosting

* + 1. Insights: For row 0 in test dataset, the probability of predictor variable belonging to class 0 is 0.54 whereas the probability of predictor variable belonging to class 1 is 0.45
  1. Accuracy:
     1. Accuracy of the logistic regression model on the training data: 87.7%
     2. Accuracy of the logistic regression model on the test data: 82.67%
     3. Insights: Accuracies of training and test data are close to each other which indicates that the model has decently generalized.
  2. Confusion matrix:

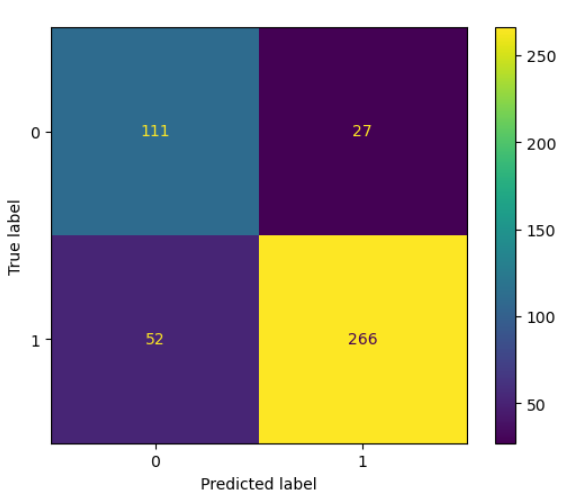
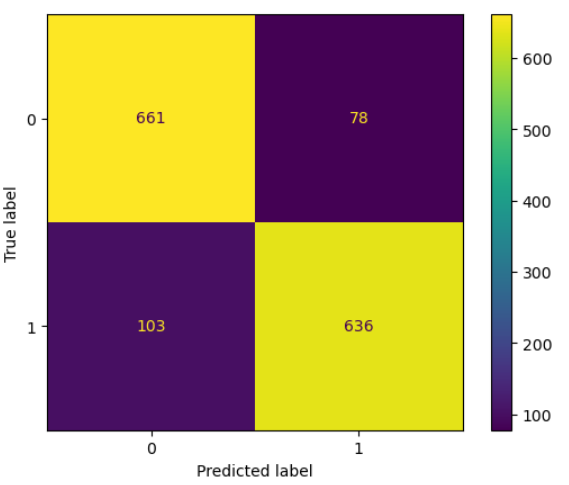


Figure 50: Confusion matrix – train and test data (Gradient Boosting)

* + 1. Insights on test data:
       1. 111 people have correctly been predicted to vote of Conservative party. 266 have correctly been predicted to vote for Labour party.
       2. 27 people have been incorrectly predicted to vote for Labour party.
       3. 52 people have been incorrectly predicted to vote for Conservative party.
  1. Classification report for training data and testing data:

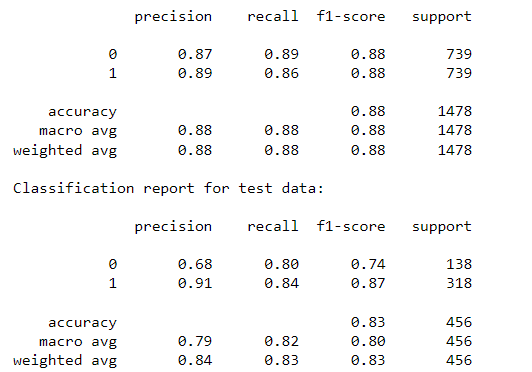


Figure 51: Classification report – training and test data (Gradient Boosting)

* + - 1. The above chart shows the precision and recall values. Row with index 0 indicates the precision recall and f-score values if 0 is considered to be positive.
      2. Row with index 1 indicates the precision, recall and f-score values if 1 is considered to be positive.
  1. ROC Curve on training data:

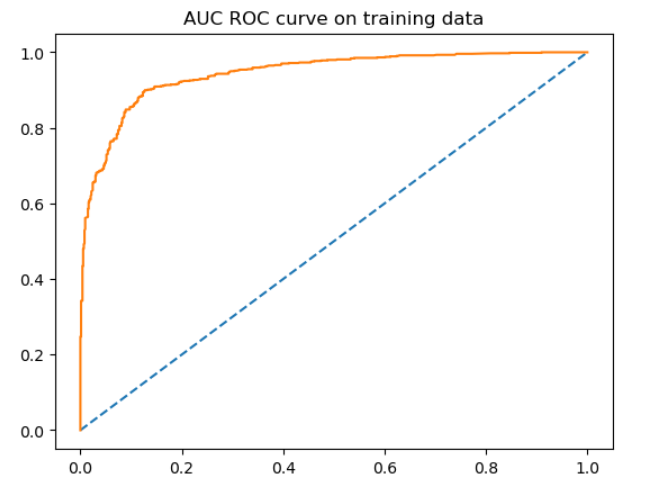


Figure 52: AUC ROC curve – training data (Gradient Boosting)

* + 1. The curve is typically obtained by plotting 1 – specificity (False Positive Rate, FPR) on the x-axis and sensitivity (True Positive Rate, TPR) on the y-axis
    2. Area under the ROC curve: AUC ROC score on training data: 0.943
  1. ROC Curve for the test data:

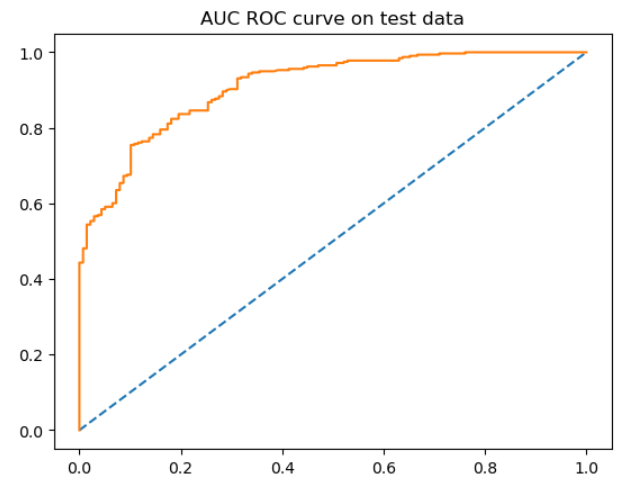


Figure 53: AUC ROC curve – testing data (Gradient Boosting)

* + 1. AUC ROC score on test data: 0.908
    2. Higher the AUC ROC score tends to 1 better is the model.

Comparison of all the models:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model Name | Accuracy | | Recall | | Precision | | F1- score | | AUC | |
| Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| Logistic Regression tuned | 81.799 | 82.894 | 0.81 | 0.83 | 0.82 | 0.91 | 0.82 | 0.87 | 0.883 | 0.917 |
| Naive Bayes tuned | 81.05 | 82.23 | 0.82 | 0.83 | 0.80 | 0.90 | 0.81 | 0.87 | 0.885 | 0.911 |
| KNN tuned | 99.93 | 78.51 | 1.00 | 0.79 | 1.00 | 0.89 | 1.00 | 0.84 | 0.999 | 0.829 |
| Bagging classifier tuned | 77.06 | 82.67 | 0.90 | 0.91 | 0.71 | 0.85 | 0.80 | 0.88 | 0.871 | 0.896 |
| Random Forest tuned | 85.79 | 84.86 | 0.85 | 0.86 | 0.86 | 0.91 | 0.86 | 0.89 | 0.934 | 0.919 |
| Adaptive boosting tuned | 84.23 | 83.33 | 0.83 | 0.85 | 0.85 | 0.90 | 0.84 | 0.88 | 0.916 | 0.905 |
| Gradient boosting tuned | 87.75 | 82.67 | 0.86 | 0.84 | 0.89 | 0.91 | 0.88 | 0.87 | 0.943 | 0.908 |

Table 11: Performance metrics of all models on elections dataset

Comparison and insights on all models:

* Out of all the models, KNN tuned algorithm, Bagging Classifier model have over learnt the training data and have performed poorly on testing data. These models have been overfitted and hence not being chosen.
* Gradient boosting has higher accuracy on training data but comparatively less accuracy on testing data. The AUC ROC scores for the data are pretty high compared to other models.
* The precision, recall and f1 score values are listed for class1 which is the Labour party. Here, precision and recall are both equally important, hence we can go with choosing a model that has the highest f1 score.

Choosing the best model:

* We choose Random Forest classifier as our final model for the following reasons:
  + The model has highest training and testing accuracy scores compared to all other models and the models are mostly equal.
  + The f1- score for Random Forest Classifier is highest than all other models.
  + The AUC ROC scores for Random Forest Classifier are the highest than all other models.

**1.8 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.**

Ans:

Business recommendations and actionable insights:

1. Feature importances of the build RandomForestClassifier are as shown below:

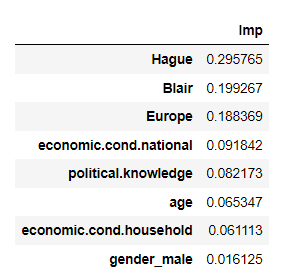


Figure 54: Feature importances of independent variables

1. The model indicates that the assessment scores given by voters to the leaders of Conservative and Labour party i.e., Blair and Hague are the most important features to predict the party the voter votes to
2. If the assessment scores given to Hague are pretty high, the voter votes to Conservative party.
3. If the assessment scores given to Blair are pretty high compared to Hague, the person votes to Labour party.
4. Euroscpetic sentiments of people play the next significant role in predicting which party they vote to. They tend to vote to leaders to support their Eurosceptic sentiment
5. Age and Gender play the least significant role in predicting voter’s choice.
6. Knowledge of the voter on current economic conditions and politics also influence their decision of which party to vote, they choose to vote to the party which best satisfies the needs for the improvement of current economic conditions.
7. Majority of the voters tend to vote to Labour party considering the current dataset.
8. Conservative party needs to focus more on supporting people’s eurosceptic sentiment and plan well for national economic growth to win voters.

**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.

2.2 Remove all the stopwords from all three speeches.

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)

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

Having a glimpse of all speeches before proceeding to the problems:

**Glimpse of Franklin D Roosevelt’s speech:**

'On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its institutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do not, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of the human spirit. The life of a man is three-score years and ten: a little more, a little less…

**Glimpse of Kennedy’s speech:**

'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God….

**Glimpse of Nixon’s speech**:

'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad…

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

**Ans:**

**Character count:**

* Total number of characters in Roosevelt's speech are 7571
* Total number of characters in Kennedy's speech are 7618
* Total number of characters in Nixon's speech are 9991
* Total number of characters in all documents are 25180

**Word count:**

* Total number of words in Roosevelt's speech are 1536
* Total number of words in Kennedy's speech are 1546
* Total number of words in Nixon's speech are 2028
* Total number of words in all documents are 5110

**Sentences count**:

* Total number of sentences in Roosevelt's speech are 68
* Total number of sentences in Kennedy's speech are 52
* Total number of sentences in Nixon's speech are 69
* Total number of sentences in all documents are 189

**2.2 Remove all the stopwords from all three speeches.**

**Removing stop words and punctuations from Roosevelt’s speech:**

* Total number of words in Roosevelt's speech before removing stopwords and punctuations are 1536
* Total number of words in Roosevelt's speech after removing stopwords and punctuations are 632
* Sample of sentence after removing stop words and punctuations in Roosevelt's speech:

‘national day inauguration since 1789 people renewed sense dedication united states washington day task people create weld together nation lincoln’

**Removing stop words and punctuations from Kennedy’s speech:**

* Total number of words in Kennedy's speech before removing stopwords and punctuations are 1546
* Total number of words in Kennedy's speech after removing stopwords and punctuations are 697
* Sample of sentence after removing stop words and punctuations in Kennedy's speech:

‘vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe’

**Removing stop words and punctuations from Nixon’s speech:**

* Total number of words in Nixon's speech before removing stopwords and punctuations are 2028
* Total number of words in Nixon's speech after removing stopwords and punctuations are 836
* Sample of sentence after removing stop words and punctuations in Nixon's speech:

‘mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together met’

**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)**

**Most frequently used words by Roosevelt:**

|  |  |
| --- | --- |
| **Word** | **Frequency** |
| Nation | 12 |
| Know | 10 |
| Spirit | 9 |
| Life | 9 |
| Democracy | 9 |
| Us | 8 |
| people | 7 |
| America | 7 |
| Years | 6 |
| Freedom | 6 |

Table 12: Most frequently used words in Roosevelt’s speech

**Most frequently used by Kennedy:**

|  |  |
| --- | --- |
| **Word** | **Frequency** |
| Let | 16 |
| Us | 12 |
| world | 8 |
| sides | 8 |
| New | 7 |
| pledge | 7 |
| Citizens | 5 |
| power | 5 |
| Shall | 5 |
| free | 5 |

Table 13: Most frequently used words in Kennedy’s speech

**Most frequently used words by Nixon:**

|  |  |
| --- | --- |
| **Word** | **Frequency** |
| Us | 26 |
| Let | 22 |
| America | 21 |
| Peace | 19 |
| world | 18 |
| New | 15 |
| nation | 11 |
| Responsibility | 11 |
| Government | 10 |
| great | 9 |

Table 14: Most frequently used words in Nixon’s speech

**2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)**

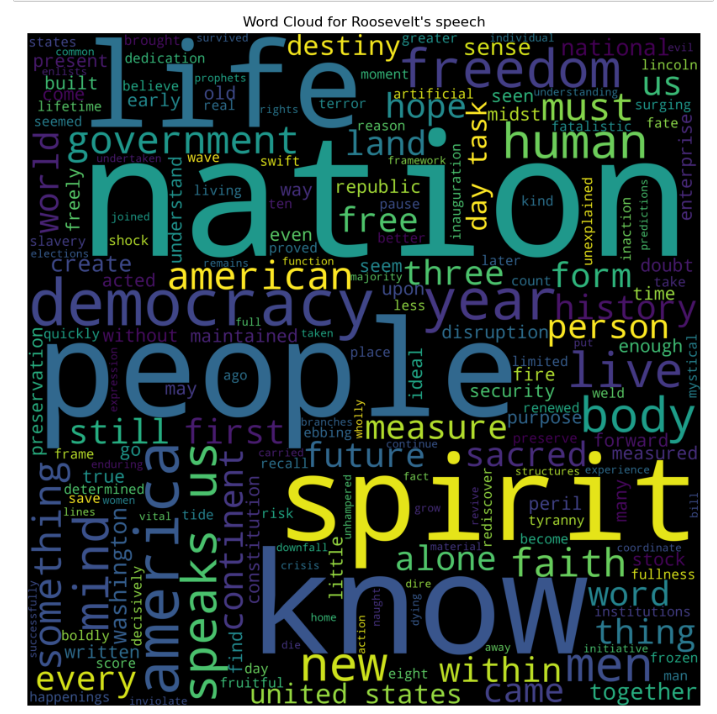
****

Figure 55: Word cloud for Roosevelt’s speech

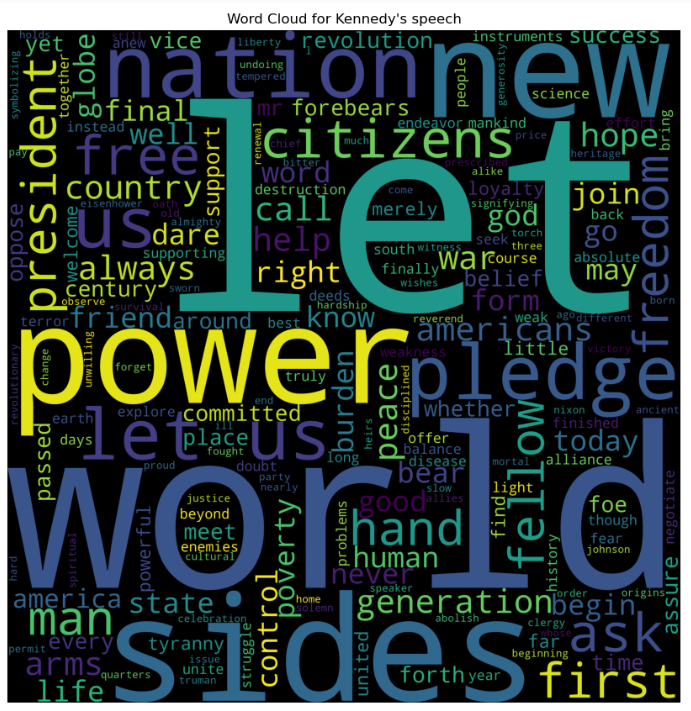
****

Figure 56: Word cloud for Kennedy’s speech

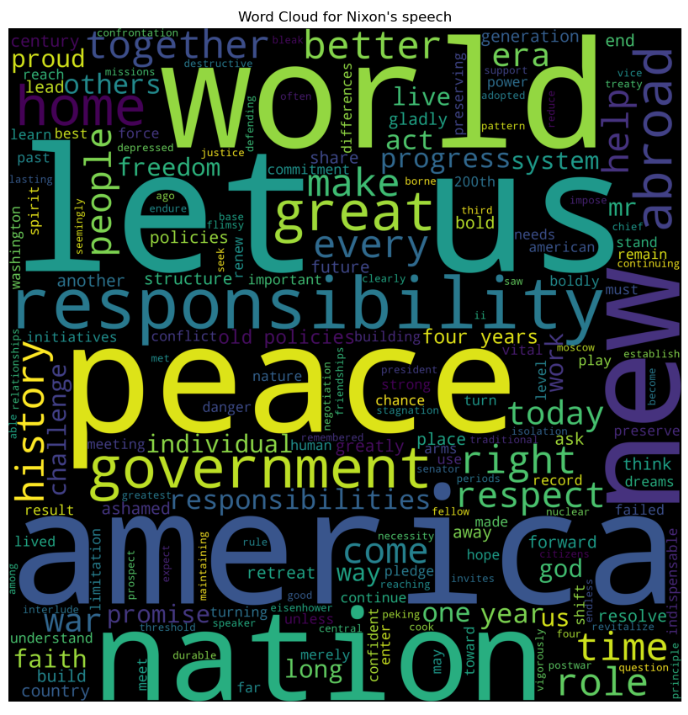
****

Figure 57: Word cloud for Nixon’s speech