

MACHINE LEARNING

Business Report

**Submitted by:**

**N. Aishwarya**

PGP-DSBA

June 2022

**Table of Contents**

**Executive summary – Machine Learning ………………………………………………………………………….….…**  **3**

**Business problem 1 – Election prediction** 3

Solution Approach 3

1.1. Descriptive statistics and inference …………………………………………………………………………………………………………… 8

1.2. Univariate, bivariate and EDA… ……………………... 19

1.3. Encoding data and modelling…. ……………………… 22

1.4. Logistic Regression and LDA. ……………………… 24

1.5. KNN Model and Naïve Bayes Model. ……………………… 37

1.6. Model Tuning, Bagging and Boosting. ……………………… 47

1.7. Performance Metrics. ……………………… 74

1.8. Key Business insights. ……………………… 78

**Business problem 2 – NLTK**…. 79

2.1. Finding the number of characters, words, and sentences for the mentioned documents………………………. 80

2.2. Remove all the stop words from all three speeches………………………. ………………… 82

2.3. Top three words occurrence ………………. 84

2.4. Plotting the word cloud of each of the speeches of the variable. ……………….. 86

**List of Figures**

|  |  |
| --- | --- |
| Figure 1: Univariate analysis of Election data | 10 |
| Figure 2: Proportion of Votes and Gender | 10 |
| Figure 3: Vote analysis of Election data | 11 |
| Figure 4: Plot of National Economic Condition with voters | 11 |
| Figure 5: National Economy Vs Vote | 12 |
| Figure 6: Household Economic Condition | 13 |
| Figure 7: Histogram plot | 15 |
| Figure 8: Outlier check | 16 |
| Figure 9: Boxplot of variable after outlier treatment | 16 |
| Figure 10: Correlation heatmap of variables | 18 |
| Figure 11: Strip plot to check how each feature affects the voting preference | 20 |
| Figure 12: Pair plot to understand the relationship between numerical values | 20 |
| Figure 13: Bar plot of Range of Data | 23 |
| Figure 14: Bar plot for Feature Importance | 25 |
| Figure 15: Confusion matrix on the training and test data | 31 |
| Figure 16: Accuracy scores for different parameters | 36 |

**List of Tables**

|  |  |
| --- | --- |
| Table 1: Description of Election data | 5 |
| Table 2: Information of Election data | 6 |
| Table 3: Summary Statistic | 7 |
| Table 4. Skewness values between independent variables | 21 |
| Table 5. Variance values between independent variables | 21 |

**List of Formulas**

|  |  |
| --- | --- |
| Formula 1: Correlation | 17 |
| Formula 2: Skewness value | 21 |
| Formula 3: Variance value | 21 |
| Formula 4: Logistic Regression | 24 |
| Formula 5: Accuracy | 26 |
| Formula 6 : LDA model | 31 |
| Formula 7 : Misclassification error | 42 |

## Executive Summary – Machine Learning:

## Machine Learning algorithms are primarily applied to “learn” from the data we provide. As additional data is provided, the model’s accuracy and efficiency to make decisions improves with subsequent training. Machine learning uses various Supervised, Unsupervised and Reinforced learning algorithms for various classification and regression problems. Identifying various trends and patterns with a huge amount of data is one of the most useful applications.

## Machine Learning’s application spans multiple industries - for example from Corporate, Defense, Politics to Education. In the corporate sector, companies apply these techniques to automate, analyze trends and patterns from the past data, predict the future, and many more. Selected applications include customer understanding, prediction of events, email spam filtering, text prediction etc.

**Business problem 1 – Election prediction**

## Problem Statement:

## 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.

## Solution Approach:

The purpose of the solutioning exercise is to explore the dataset using machine learning techniques to arrive at the customer segmentation, thus enabling business strategies customized to them. Below is the data dictionary for given problem:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **Vote** | Party choice: Conservative or Labour |
| **Age** | Voter’s age in years |
| **Economic.cond.national** | Assessment of current national economic conditions, 1 to 5. |
| **Economic.cond.household** | Assessment of current household economic conditions, 1 to 5. |
| **Blair** | Assessment of the Labour leader, 1 to 5. |
| **Hague** | Assessment of the Conservative leader, 1 to 5. |
| **Europe** | An 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment. |
| **Political.knowledge** | Knowledge of parties' positions on European integration, 0 to 3. |
| **Gender** | Voter’s gender - Female or male. |

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

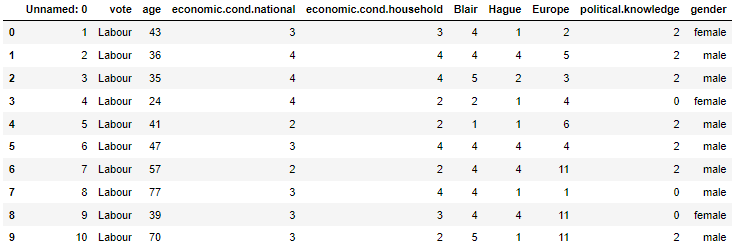
**Solution:**

Firstly, import the necessary libraries required for the problem in the Jupiter Notebook file and

run them. Read the “Election\_data.csv” file for EDA.

Since the ‘Vote’ variable is the target variable, we therefore have ‘vote’ as the dependent and rest 8 variables as the independent variables.

**Below are the top few records to get a feel of the data structure**

****

*Table 1: Description of Election data*

### We will drop the first column ‘Unnamed: 0’ column as this is not important for our study

### Checking for duplicate values:

Number of duplicate rows = 8

**Observation**:

* There are 8 duplicate values.

### We will drop the duplicate rows and proceed with EDA process.

### Checking the data (after dropping the unnamed column & duplicate values):

### 

### List of Columns:

### 

### Let us understand the number of rows and columns in the dataset.

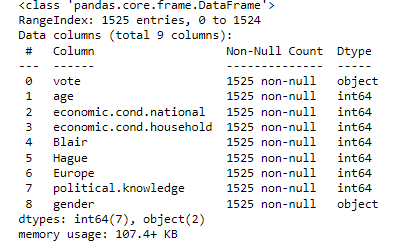
### Checking the shape of the data:

**Number of rows:** 1,517

**Number of columns:** 9

Let us check the basic info about the dataset and also the statistical summary.

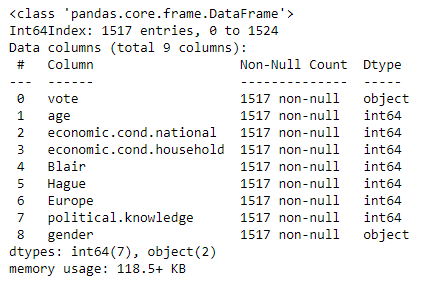
### Retrieving the list of fields along with their data type:

****

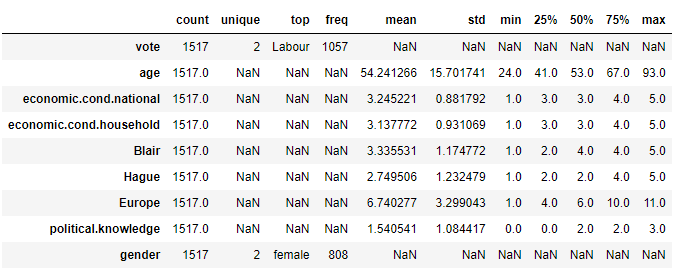
*Table 2: Information of Election data*

Unnamed column has been dropped from the dataframe. All the independent continuous columns have an integer datatype although some of the category columns have object datatype and that can be handled using one hot coding. The dependent column has object datatype.

**Checking the info () function again:**



**Checking the Summary Statistic:**

****

*Table 3: Summary Statistic*

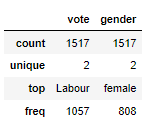
**Checking for Missing Values:**

### 

**Observations:**

* There are no missing values.

**Getting the summary statistics of the object variable:**

****

### Observations:

* The data set contains 1525 rows, 10 columns.
* In the given data set, there are 2 Object type features: vote and gender, 7 integer type features, where 'VOTE' is the target variable and all other are predictor variable.
* The first column is an index ("Unnamed: 0") as this only serial no, we can remove it.
* There are no null values present in the dataset.
* There are 8 duplicate values. We have dropped it since it’s of no use.
* Every variable has no missing or non-numerical values present.
* Almost all the variables have equal mean and median values.
* The minimum age of the voters is 24 years and maximum are 93 years.
* Both labour and conservative parties are equally distributed.
* The male and female voters are briefly divided as “Labour” and “Conservative” parties. People prefer Labour party more over the conservative party.
* The number of female voters is more than male voters.

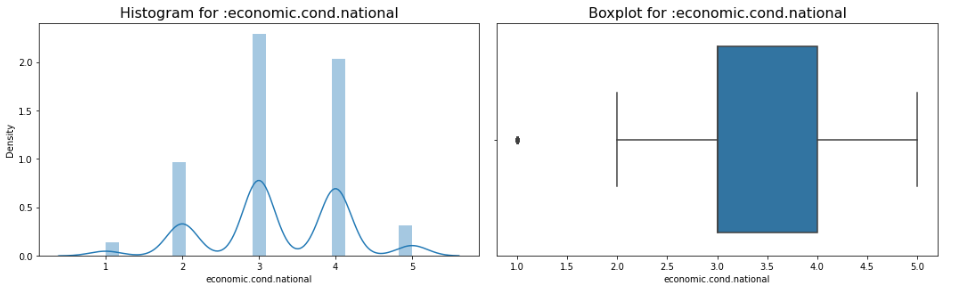
# Analysis 1.2: Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

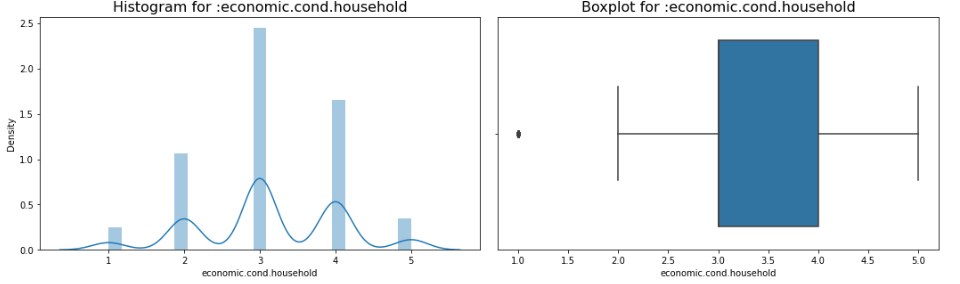
### 

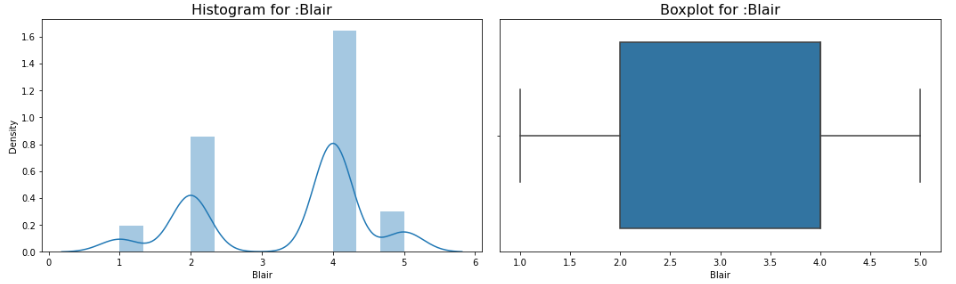
### Solution: Let us clearly look onto the distribution of the variables with outliers in detail by plotting histogram and boxplot simultaneously.

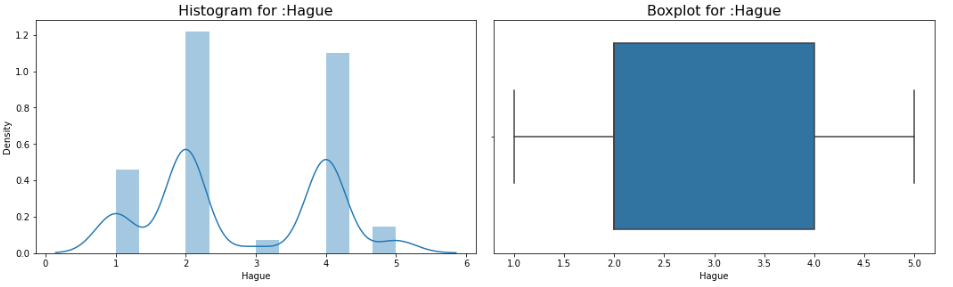
### Univariate Analysis:

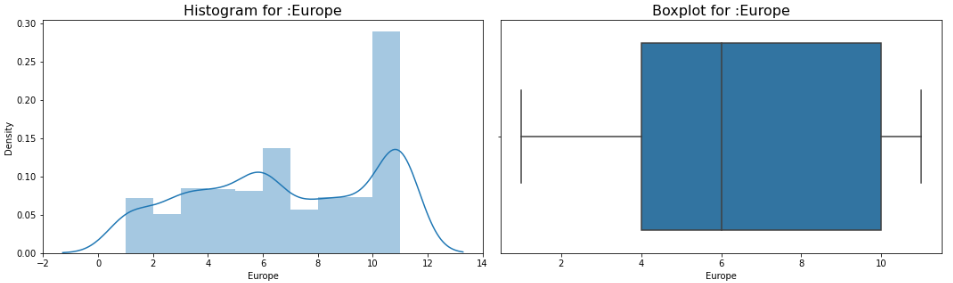
### 

**

**

**

**

**

### 

*Figure 1: Univariate analysis of Election data*

### Proportion of Votes and Gender:

### 

### 

*Figure 2: Proportion of Votes and Gender*

### 

### Observations:

### Distribution of Male and female voters across the age is almost in equal proportion with Females slightly higher in number in comparison to Males.

### 70% of the voters preferred Labour parties and only 30% of the voter preferred conservative parties.

### 

### Plot of Vote and Gender:

### 

*Figure 3: Vote analysis of Election data*

**Observations:**

* Female voter prefers Labour parties more than Conservative parties.
* Male Labour voters are more than Male Conservative voters. i.e., Voter prefers to voter Labour voter as compared to Conservative voters.

**Average mean age of Labour is: 53.1**

**Average mean age of Conservative is: 56.8**

### National Economic Condition with voters:

### 

### 

*Figure 4: Plot of National Economic Condition with voters*

### Plot of National Economy Vs Vote:

### 

*Figure 5: National Economy Vs Vote*

### Household Economic Condition:

### 

### 

### 

*Figure 6: Household Economic Condition*

### Plot of Household Economy with Vote:

### 

### Tom Blair vs William Hague:

### 

### Let us define a function 'univariateAnalysis\_category' to display information as part of univariate analysis of categorical variables.

The function should display the frequency of all the levels within the field and display a frequency plot.

### Getting unique counts of Categorical Variables

### 

### 

### 

### 

### Histogram Plot:

### 

*Figure 7: Histogram plot*

### Outlier’s check:

### 

*Figure 8: Outlier check*

### Now let’s remove the outlier from the dataset by performing IQR method. Below figure shows the boxplot of variable after outlier treatment.

### 

*Figure 9: Boxplot of variable after outlier treatment*

### Bivariate Analysis and Multivariate Analysis:

### Correlation between variables of the dataset:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related.

**Formula 1: Correlation**

Where,

Covariance of and

Standard deviation of

Standard deviation of

### Let us check the correlation of the variables:

### 

### CORRELATION HEATMAP :

### 

*Figure 10: Correlation heatmap of variables*

### Observations:

### We can see that most of the variables are negatively correlated and there is less chance of multicollinearity between the variables.

### Age variable is negatively correlated with political knowledge and household economic condition i.e., voters age does not depend on the political knowledge of the parties.

### We will use Strip plot to check how each feature affects the voting preference.

### 

### 

### 

*Figure 11: Strip plot to check how each feature affects the voting preference*

### Pair Plot:

The Pair Plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns. The pair plot function in seaborn makes it very easy to generate joint scatter plots for all the columns in the data.

### 

*Figure 12: Pair plot to understand the relationship between numerical values*

### Here, Red represents “Labour” party’s and blue represents “Conservative” parties.

### SKEWNESS VALUE:

### Formula 2:

### Skewness = 3 \* (Mean – Median) / Standard Deviation.

|  |  |
| --- | --- |
| Variables | Values |
| Hague | 0.146 |
| age | 0.139 |
| economic. cond. household | 0.0918 |
| economic. cond. national | -0.069 |
| Europe | -0.142 |
| political. knowledge | -0.423 |
| Blair | -0.539 |

### *Table 4. Skewness values between independent variables*

### Inferences:

### Almost all the variables have skewness value close to zero.

### Except “Age” and “Hague”, all other variables are negatively skewed.

### We will check the independent variables variance and check for variables having no variance or almost zero variance(variance < 0.1). They will be having almost no influence on the classification.

### VARIANCE VALUE:

### Formula 3:

|  |  |
| --- | --- |
| Variables | Values |
| age | 246.545 |
| economic. cond. national | 0.729 |
| economic. cond. household | 0.785 |
| Blair | 1.380 |
| Hague | 1.519 |
| Europe | 10.884 |
| political. knowledge | 1.176 |

### *Table 5. Variance values between independent variables*

### The inferences drawn from the above Exploratory Data analysis:

### We can observe that relatively younger people have voted for “Labour “ party in comparison to that of older people who voted for “Conservative” party.

### There is an evenly distributed number of people when it comes to their knowledge about their party’s position on European integration.

### Majority of European people have voted for “Labour” party.

### There exists an outlier for economic.cond.national and economic.cond.household variable.

### The variables “Age” and “Hague” have high skewness values compared to other variables.

### We can observe that the variable “Age” has high variance value. This implies that the age variable affects the voters preference.

### Age variable has highly skewed and normally distributed.

### “Economic.cond.national” and “Economic.cond.household” is normally distributed and right skewed.

### “Blair” and “Europe” variable is left skewed.

### Analysis 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).

### Solution: Drop First is used to ensure that multiple columns created based on the levels of categorical variable are not included else it will result in to multicollinearity . This is done to ensure that we do not land in to dummy trap.

### Adding a new column category for the age group:

|  |  |
| --- | --- |
| 1 | 199 |
| 2 | 475 |
| 3 | 415 |
| 4 | 366 |
| 5 | 62 |

### We can observe that “vote” has codes, 0 represent “conservative” and 1 represent “Labour” votes. “Gender” variable is also assigned with respective codes 0s and 1s.

### Getting dummies of the variables:

### 

### 

### Bar plot of Range of Data:

### 

*Figure 13: Bar plot of Range of Data*

### Scaling:

### Scaling is necessary only for model which is based on distance rule. In this problem, let us perform scaling only for KNN modelling.

### Let us split the dataset into 70% training and 30% test size.

### Train Test Split:

### Checking the dimensions of the training and test data:

Dimensions of the training and test data:

X\_train (1061, 41)

X\_test (456, 41)

y\_train (1061,41)

y\_test (456,41)

Total Observations are 1517

**Y train value count**:

|  |  |
| --- | --- |
| 1 | 0.697 |
| 0 | 0.303 |

**Y test value count:**

|  |  |
| --- | --- |
| 1 | 0.697 |
| 0 | 0.303 |

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

### Solution: Logistic Regression Model:

**Formula 4: Logistic Regression**

**, Where**

### Fitting a Logistic Regression Model:

### 

**Bar plot for Feature Importance:**

****

*Figure 14: Bar plot for Feature Importance*

### Getting the Predicted Classes and Probs:

### 

### Model Evaluation for training data:

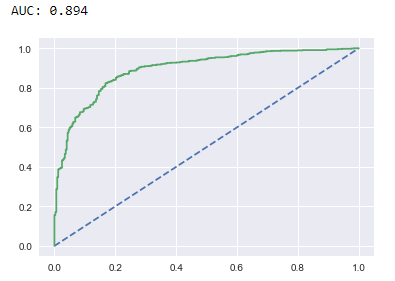
logit\_train\_accuracy: 0.85

logit\_train\_precision: 0.77

logit\_train\_recall: 0.71

logit\_train\_f1: 0.74

### AUC and ROC for the training data:



### Model Accuracy Scores :

### Formula 5: Accuracy

### Model Evaluation for testing data:

logit\_test\_accuracy: 0.84

logit\_test\_precision: 0.76

logit\_test\_recall: 0.67

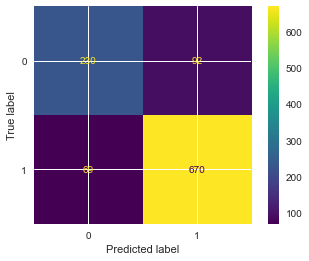
logit\_test\_f1: 0.71

### AUC and ROC for the test data:

### 

### Confusion Matrix for the training data





# Classification Report for training data:

# 

### Confusion Matrix for test data:

### 

### 

# Classification Report for testing data:

### 

# ROC-AUC Curve of Logit Model:

# 

# ROC - Logistic Regression Train Data:

# 

# ROC - Logistic Regression Test Data:

# 

### Applying GridSearchCV for Logistic Regression:

### Fitting GridSearchCV Model:

### 

# GridSearchCV best Parameters and Estimators:

# 

# Getting the probabilities on the test set:

# 

# 

# Confusion matrix on the training data:

# 

# Confusion matrix on the test data:

# 

*Figure 15: Confusion matrix on the training and test data*

### LDA Model:

### Formula 6 : LDA model

### The LDA model gives linear combinations of the predictor variables as follows:

### 

### Where: DS = Discriminant Score

### β's = Discriminant weight (coefficients)

### X’s = Explanatory (Predictor or independent) variables

### Prediction:

### 

# LDA Model Accuracy (for training data):

LDA\_train\_accuracy: 0.84

LDA\_train\_precision: 0.75

LDA\_train\_recall: 0.72

LDA\_train\_f1: 0.73

# Classification report of Train Data:

# 

# LDA Model Accuracy (for testing data):

LDA\_test\_accuracy: 0.85

LDA\_test\_precision: 0.77

LDA\_test\_recall: 0.72

LDA\_test\_f1: 0.74

# Classification report of Test Data:

# 

# Confusion matrix on the training data:

# 

# 

# Confusion matrix on the test data:

# 

# 

# ROC-AUC Curve of LDA Model:

# 

# ROC - LDA Train Data:

# 

# ROC - LDA test Data:

# 

### We Will change the cut-off values for maximum accuracy.

### 

# 

# 

# 

# 

*Figure 16: Accuracy scores for different parameters*

We see that 0.5 and 0.6 gives better accuracy than the rest of the custom cut-off values. But 0.5 cut-off gives us the best 'f1-score'. Here, we will take the cut-off as 0.5 to get the optimum 'f1' score. Let us evaluate the predictions of the test data using these cut-off values.

# Confusion matrix and classification report on predicted cut-off data set:

# 

# 

# 

### Inferences:

# From the above model accuracies, we can see that LDA looks better compares to logistic regression with a small difference in the training accuracies. Both Logistic regression and LDA does not requires scaling because they are not affected by scaled values.

# Analysis 1.5: Apply KNN Model and Naïve Bayes Model. Interpret the results.

#### 

### Solution: Fitting KNN Model:

### 

### Parameters of KNN Model:

### 

### KNN Model Accuracy(for training data):

KNN\_train\_accuracy: 0.84

KNN\_train\_precision: 0.77

KNN\_train\_recall: 0.68

KNN\_train\_f1: 0.72

### Classification report of Train Data:

### 

### KNN Model Accuracy(for testing data):

KNN\_test\_accuracy: 0.82

KNN\_test\_precision: 0.72

KNN\_test\_recall: 0.67

KNN\_test\_f1: 0.7

### Classification report of Test Data:

### 

### Confusion matrix on the training data:

### 

### 

### Confusion matrix on the test data:

### 

### 

### ROC - KNN Train Data:

### 

### ROC - KNN test Data:

### 

### ROC-AUC Curve of KNN Model:

### 

### Default value n\_neighbors=5, let’s check the performance for K=7

### 

### Performance Matrix on train data set:

### 

### Performance Matrix on test data set:

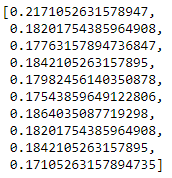
### 

Run the KNN with no of neighbours to be 1,3, 5.19 and \*Find the optimal number of neighbours from K=1,3,5,7....19 using the Mis classification error

### Formula 7:

**Misclassification error (MCE) = 1 - Test accuracy score**.

Calculated MCE for each model with neighbours = 1,3,5...19 and find the lowest MCE.



### Plot misclassification error vs k (with k value on X-axis) using matplotlib:

### 

### 

### For K = 11 it is giving the best test accuracy let’s check train and test for K=13 with other evaluation metrics.

### 

### Performance Matrix on train data set:

### 

### Performance Matrix on test data set:

### 

# As the difference between train and test accuracies is 3.68 % which is less than 10%(Industry standard).So, it is a valid model.

### Gaussian Naive Bayes:

### Fitting Naïve Bayes Model:

### 

# Parameters of Naïve Bayes Model:

# 

# Predicted value on test data:

# 

# MNB Model Accuracy( for training data):

MNB\_train\_accuracy: 0.83

MNB\_train\_precision: 0.72

MNB\_train\_recall: 0.71

MNB\_train\_f1: 0.71

# Classification report of Train Data:

# 

# MNB Model Accuracy(for testing data):

MNB\_test\_accuracy: 0.84

MNB\_test\_precision: 0.74

MNB\_test\_recall: 0.72

MNB\_test\_f1: 0.73

# Classification report of Test Data:

# 

# Confusion matrix on the training data:

# 

# 

# Confusion matrix on the test data:

# 

# 

# ROC - MNB Train Data:

# 

# ROC - MNB test Data:

# 

# ROC-AUC Curve of MNB Model:

# 

# Looking at Recalls, Training accuracy and Test accuracy. Model seems to be performing well.

# Inferences:

# By observing the classification report of Naïve bayes testing classification report, we can see that we have received a better f1-score on Labour voters. There is a 1% increase in the prediction for Labour voters with naïve bayes model while 73% of f1-score for conservative voters.

# As of now, this model looks good since there is only 2% difference with the testing and training model accuracies. We can see that naïve bayes is having less overfitting issue. To an extent, we have received a quiet good precision and recall rates.

# Analysis 1.6: Model Tuning, Bagging (Random Forest should be applied for Bagging), and boosting.

### Solution: Let us use GRIDSEARCHCV for hyper tuning the models. Before applying the bagging classifier, let us use random forest and apply bagging base estimator. Let us apply gradient boosting and Ada boost classifier as boosting algorithm. Naïve bayes does not hold with a greater number of parameters. So, it is difficult to tune the naïve bayes algorithm. We will be tuning all the base models with different combination of parameters.

### Model Tuning:

### 

### Linear Regression with SMOTE

### Model Accuracy:

logit\_res\_test\_accuracy: 0.85

logit\_res\_test\_precision: 0.79

logit\_res\_test\_recall: 0.69

logit\_res\_test\_f1: 0.74

### Classification Report:

### 

### 

### LDA with SMOTE:

### Model Accuracy:

LDA\_res\_test\_accuracy: 0.7

LDA\_res\_test\_precision: 0.0

LDA\_res\_test\_recall: 0.0

LDA\_res\_test\_f1: 0.0

**Classification Report:**

### 

### KNN with SMOTE

### Model Accuracy:

KNN\_res\_test\_accuracy: 0.82

KNN\_res\_test\_precision: 0.74

KNN\_res\_test\_recall: 0.64

KNN\_res\_test\_f1: 0.69

### Classification Report:

### 

### SVM with SMOTE

### Model Accuracy:

SVM\_res\_test\_accuracy: 0.7

SVM\_res\_test\_precision: 0.0

SVM\_res\_test\_recall: 0.0

SVM\_res\_test\_f1: 0.0

### 

### Classification Report:

### 

### MNB with SMOTE

### 

### Model Accuracy:

MNB\_res\_test\_accuracy: 0.74

MNB\_res\_test\_precision: 0.61

MNB\_res\_test\_recall: 0.41

MNB\_res\_test\_f1: 0.49

### Classification Report:

### 

### Hyperparameter tuning using GridSearchCV:

### Logistic Regression with GridSearchCV:

### 

### 

### 

### Model Accuracy(for training data):

logit\_train\_accuracy: 0.83

logit\_train\_precision: 0.68

logit\_train\_recall: 0.83

logit\_train\_f1: 0.75

### Classification Report for training data:

### 

### Model Accuracy (for testing data):

logit\_test\_accuracy: 0.84

logit\_test\_precision: 0.76

logit\_test\_recall: 0.67

logit\_test\_f1: 0.71

### Classification Report for testing data:

### 

### Confusion matrix on the training data:

### 

### 

### Confusion matrix on the testing data:

### 

### 

### ROC - Logistic Regression Train Data:

### 

### ROC - Logistic Regression Test Data:

### 

### ROC-AUC Curve of Logistic Regression:

### 

### Inferences:

### There is no much difference in the tuned and base model for logistic regression. We have received 87% of f1-score on “Labour” votes and 74% of better prediction for “Conservative” voters.

### 

### Linear Discriminant Analysis with GridSearchCV:

### 

### 

### Model Accuracy for training data:

LDA\_train\_accuracy: 0.84

LDA\_train\_precision: 0.75

LDA\_train\_recall: 0.72

LDA\_train\_f1: 0.73

### Classification Report:

### 

### Model Accuracy for testing data:

LDA\_test\_accuracy: 0.85

LDA\_test\_precision: 0.77

LDA\_test\_recall: 0.72

LDA\_test\_f1: 0.74

### Classification Report:

### 

### Confusion Matrix for LDA Training set:

### 

### 

### Confusion Matrix for LDA Test set:

### 

### 

### ROC - LDA Train Data:

### 

### ROC - LDA Test Data:

### 

### ROC-AUC Curve of LDA Model:

### 

### Inferences:

### The base model and tuned model are having no differences. We can see that 89% better prediction is for Labour voters and 74% F1-score is conservative voters are obtained from LDA models. Conservative voters are correctly and more accurately predicted in LDA model compared to Logistics regression model.

### KNN Model with GridSearchCV:

### 

### 

### KNN Model Accuracy for training data:

KNN\_train\_accuracy: 0.81

KNN\_train\_precision: 0.72

KNN\_train\_recall: 0.64

KNN\_train\_f1: 0.67

### Classification Report for training data:

### 

**KNN Model Accuracy for Testing data:**

KNN\_test\_accuracy: 0.83

KNN\_test\_precision: 0.75

KNN\_test\_recall: 0.67

KNN\_test\_f1: 0.7

### Classification Report for testing data:

### 

### Confusion Matrix for KNN Training set:

### 

### 

### Confusion Matrix for KNN Test set:

### ,

### 

### ROC - KNN Train Data:

### 

### ROC - KNN test Data:

### 

### ROC-AUC Curve of KNN Model:

### 

# Inferences:

# Once tuned, we can see that after tuning the F1-score for Labour voters have increased by 1% on testing report. This model looks perfect since it has better f1-score, precision, recall and model accuracies comparing to rest of the models. Let us compare in detail using a tabular format. Before that, let us also check for bagging and boosting classifier models reports.

# From the above model accuracies, KNN model have more accuracy on training dataset compares to all other model. All model has overfit problems. Let us consider the modes with which less than 10% difference in training and testing as a standard model. So, let us perform the prediction with all models and check their precision, recall, f1-score and model accuracy. We will use F1- score to compare the model efficiency.

### Support Vector Machine with GridSearchCV:

### 

### 

### SVM Model Accuracy for training data:

SVM\_train\_accuracy: 0.82

SVM\_train\_precision: 0.65

SVM\_train\_recall: 0.84

SVM\_train\_f1: 0.73

### Classification Report for training data:

### 

### SVM Model Accuracy for testing data:

SVM\_test\_accuracy: 0.82

SVM\_test\_precision: 0.65

SVM\_test\_recall: 0.86

SVM\_test\_f1: 0.74

### Classification Report for testing data:

### 

### Confusion Matrix for SVM model Training set:

### 

### 

### Confusion Matrix for SVM model Test set:

### 

### 

### ROC - SVM Train Data:

### 

### ROC - SVM test Data:

### 

### ROC-AUC Curve of SVM\_model:

### 

### Bagging using Random Forest:

### For Bagging, let us import random forest classifier from sklearnmodel library. Apply all the parameters and fit the model.

### 

### Model Accuracy for training data:

Bagging\_train\_accuracy: 0.88

Bagging\_train\_precision: 0.76

Bagging\_train\_recall: 0.87

Bagging\_train\_f1: 0.81

### Classification Report for training data:

### 

### Model Accuracy for testing data:

Bagging\_test\_accuracy: 0.83

Bagging\_test\_precision: 0.7

Bagging\_test\_recall: 0.78

Bagging\_test\_f1: 0.74

### Classification Report for testing data:

### 

### Confusion Matrix for Bagging Model Training set:

### 

### 

### Confusion Matrix for Bagging Model Test set:

### 

### 

### ROC - Bagging Train Data:

### 

### ROC - Bagging test Data:

### 

### ROC-AUC Curve of Bagging Model:

### 

### Inferences:

### After applying the bagging classifier, we have received 88% of model accuracy with better F1-score. But once we check for the performance of testing model, we have only 83% model accuracy.

### We can see that F1-score for conservative and Labour voters predicted has drop down drastically by comparing with other models. We can see there is a 6% difference in the training and testing model accuracies of bagging classifier.

### Boosting:

### We are using two boosting techniques to find out the best model, Gradient boosting and Ada boosting classifier.

### XGBoost:

### 

### Model Accuracy for training data:

### 

XGB\_train\_accuracy: 0.88

XGB\_train\_precision: 0.83

XGB\_train\_recall: 0.77

XGB\_train\_f1: 0.8

### Classification Report for training data:

### 

### Model Accuracy for testing data:

XGB\_test\_accuracy: 0.84

XGB\_test\_precision: 0.76

XGB\_test\_recall: 0.67

XGB\_test\_f1: 0.71

### Classification Report for testing data:

### 

### Confusion Matrix for XGB Model Training set:

### 

### 

### Confusion Matrix for XGB Model Test set:

### 

### 

### ROC - XGB Train Data:

### 

### ROC - XGB test Data:

### 

### ROC-AUC Curve of XGB Model:

### 

### Gradient Boosting Classifier:

### 

**Model Accuracy for training set:**

GBC\_train\_accuracy: 0.99

GBC\_train\_precision: 0.98

GBC\_train\_recall: 0.99

GBC\_train\_f1: 0.98

**Model Accuracy for testing set:**

GBC\_test\_accuracy: 0.8

GBC\_test\_precision: 0.68

GBC\_test\_recall: 0.64

GBC\_test\_f1: 0.66

### Ada Boost:

### 

### Classification Report for training set:

### 

### Classification Report for testing set:

### 

### AUC \_ROC Curve Boosting Train:

### 

### AUC \_ROC Curve Boosting Test:

### 

### Inferences:

### 

# Analysis 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.

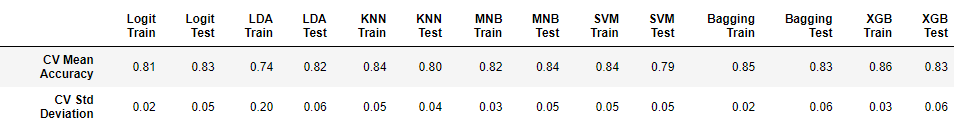
# 

# Solution: Performance Metrics for training and testing set is:

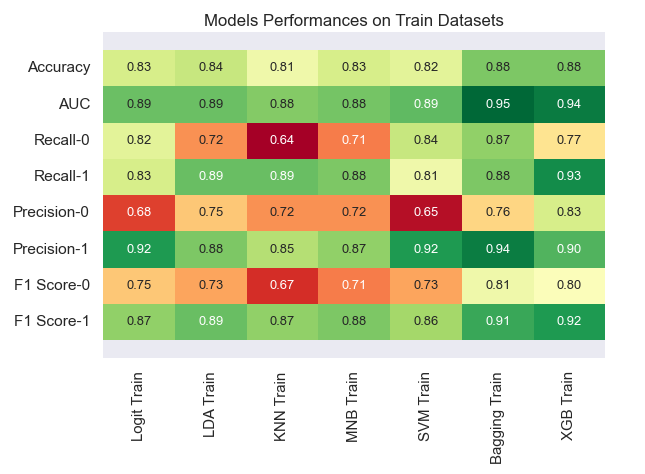
# 

# 

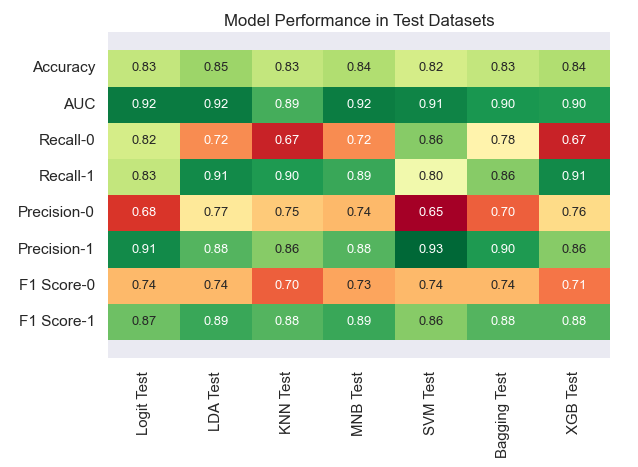
### Mean Accuracy and standard deviation:



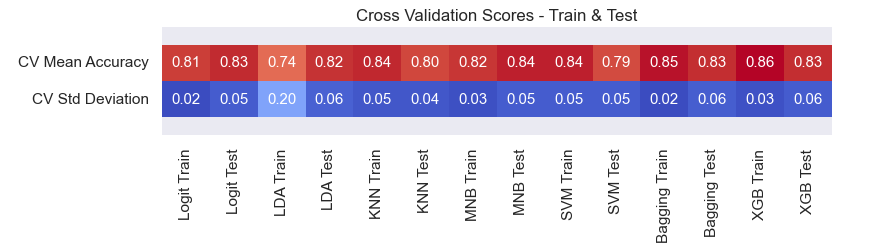
**Models Performances on Train Datasets:**

****

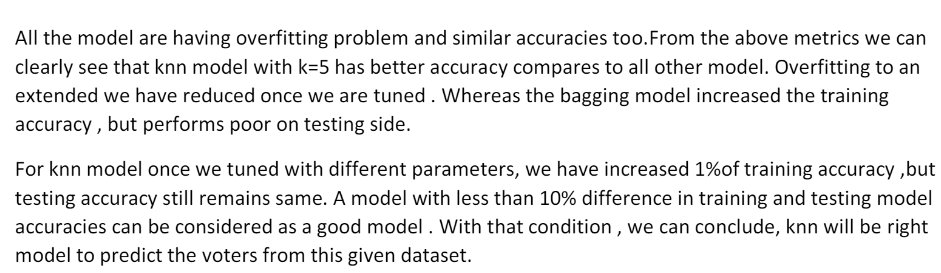
**Model Performance in Test Datasets:**

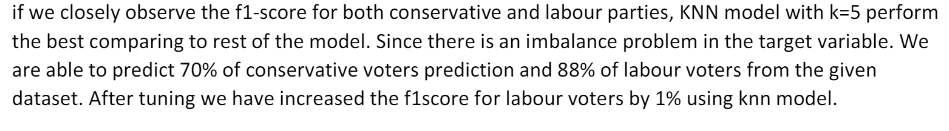


**Cross Validation Scores:**

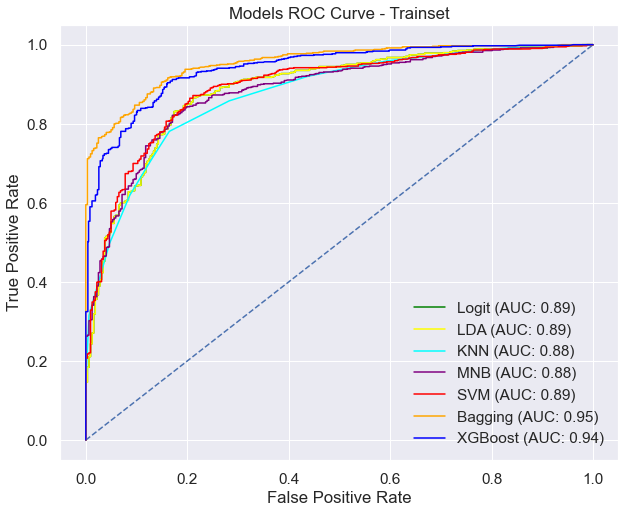


**Inferences:**

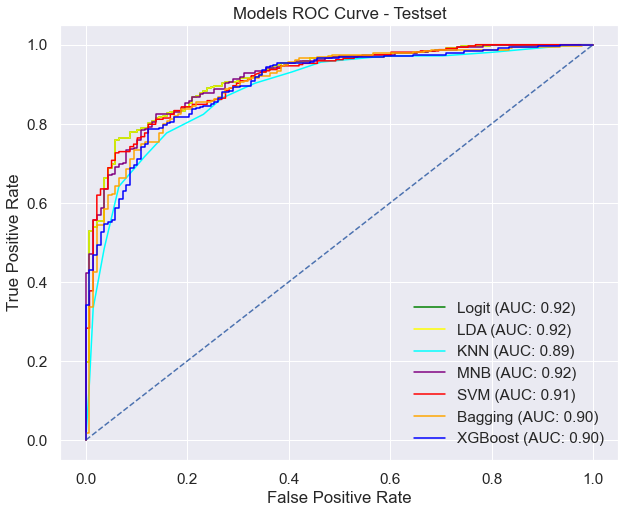




**Models ROC Curve – Trainset:**



**Models ROC Curve – Test set:**



# Analysis 1.8: Based on these predictions, what are the insights?

**There are multiple business insights that have been generated through EDA and the model building exercise**. Below insights will help to create an exit poll to predict overall win and seats covered:

### Labour party is better poised to win the elections, in comparison to Conservative party

### The male and female voters are briefly divided as “Labour” and “Conservative” parties. People prefer Labour party more over the conservative party

### Relatively younger people shall vote for “Labour “ party in comparison to that of older people who will vote for “Conservative” party.

### There is an evenly distributed number of people when it comes to their knowledge about their party’s position on European integration.

### Majority of European people are likely to vote for “Labour” party.

### There exists an outlier for economic.cond.national and economic.cond.household variable.

### We can observe that the variable “Age” has high variance value. This implies that the age variable affects the voters preference.

### Business Problem 2: NLTK

## Problem Statement:

## 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:

## President Franklin D. Roosevelt in 1941

## President John F. Kennedy in 1961

## President Richard Nixon in 1973

## Solution Approach:

## Inaugural File ids:

## 

## Length of Inaugural File id is: 59

## Inaugural speech:

## 

## Snapshot of Words in '1941-Roosevelt.txt':

## 

## Snapshot of Words in '1961-Kennedy.txt':

## 

## Snapshot of Words in '1973-Nixon.txt':

## 

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

### Solution: Number of Words:

### 

# 

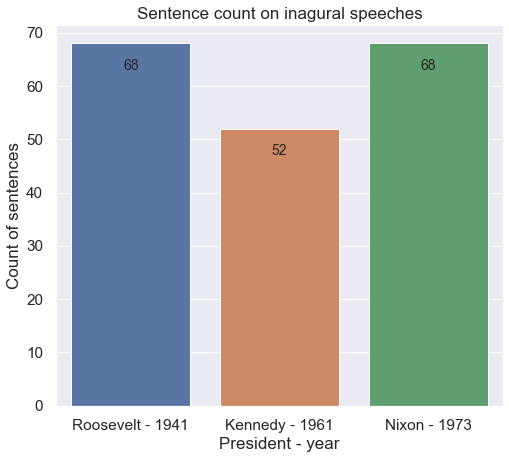
### Number of characters:

### 

### 

### Number of sentences:

### 



# Analysis 2.2: Remove all the stop words from all three speeches

# Solution:

# Lower case conversion:

# 

# Remove punctuation:

# 

# 

# Speech of president Roosevelt without stop words:

# 

# Speech of president Kennedy without stop words:

# 

# Speech of president Nixon without stop words:

# 

# Analysis 2.3: Which word occurs the greatest number of times in his inaugural address for each president? Mention the top three words.

# Solution:

# Inaugural speech of Roosevelt - 1941:

# The most commonly occurred words are:

# 

# 

# Inaugural speech of Kennedy – 1961:

# The most occurred words are:

# 

# Inaugural speech of Nixon – 1973:

# The most occurred words are:

# 

# Analysis 2.4: Plot the word cloud of each of the speeches of the variable.

# Solution:

# Word Cloud for Roosevelt after cleaning:

# 

# 

# Word Cloud for Kennedy after cleaning:

# 

# Word Cloud for Nixon after cleaning:

# 