Machine Learning Project

S.Vishal

PGPDSBA JULY

**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.

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

We have read the dataset Election\_data.xlsx.

Let’s see the head of the data :

Application

Description automatically generated with low confidence

We found that there are 2 categorical variables.

Description of dataset

Graphical user interface

Description automatically generated with low confidence

We can see that most values are ranging between 0 to 11 except Age so to put it in the same range we will try binning in subsequent steps.

• Checking for null values – From panda’s null check value function, we found below results which say that there are no null values in the dataset.

Graphical user interface

Description automatically generated with medium confidence

• Checking Duplicate records – From Pandas duplicate function we found 8 duplicate values so we have dropped them.

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

We have done Null Value check and found that there are No Null Values in the Dataset.

• Shape of the dataset –



• Info of the dataset –

Text

Description automatically generated

Data set contains 2 Categorical columns Vote and gender, all other columns are of Integer type.

• Unique value counts for all the object data types –

Text, letter

Description automatically generated

• EDA

**Univariate analysis**

How many votes did each party get? (shown using countplot)

Labour 1057

Conservative 460

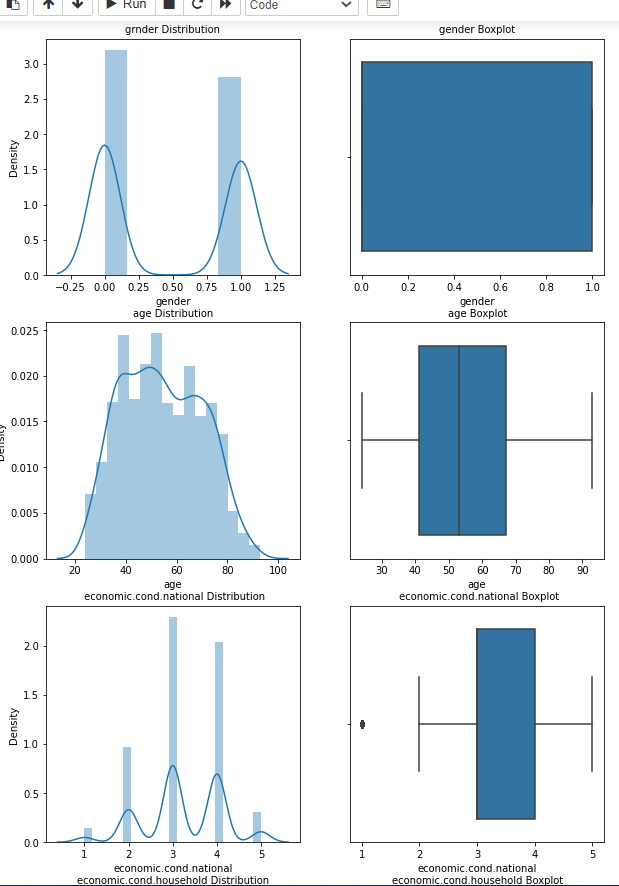
Chart, bar chart

Description automatically generated

Figure 1: Party vs Votes

Labour party have Got 1057 votes & Conservative party have got 460 votes.

Density distribution using distribution plots & Boxplots



Shape, square

Description automatically generated

Figure 2: Distribution plot for election data

As we can see from the plot, most of the data points are negatively/left skewed except Age & Hague.

Skewness is given below –

Graphical user interface, text, application

Description automatically generated

Also from the boxplot, we find that there are outliers present on economic condition national & household attributes, but on checking further we see that the ratings can actually be 1 for certain cases (Assessment of the economic conditions being a scale of 1-5). So all the values are correct and it will not be a wise decision to treat these outliers.

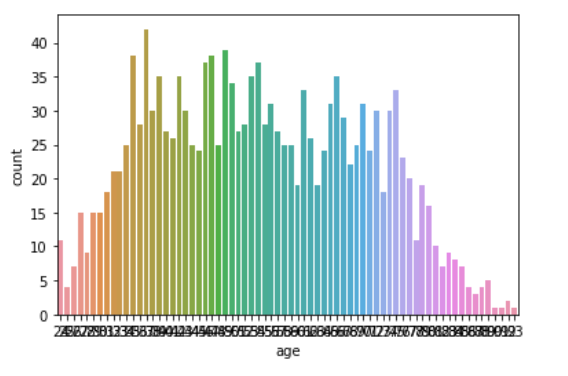


Figure 3:Age of voters

Most of the voters fall in the age bracket of 40s and 50s years.

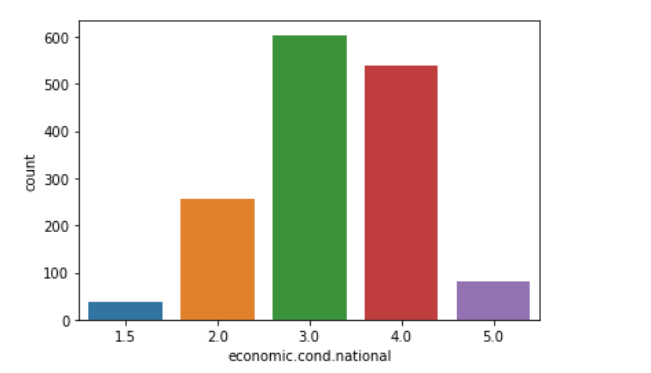


Figure 4:Current National Economic conditions

Most frequent rating is 3 and least frequent is 1. We can say that the average rating lies between 2-3.

Chart, bar chart

Description automatically generated

Figure 5:Current Household Economic Condition

Most frequent rating is 3 and least frequent rating is 1. We can say that the average rating lies between 2-3.

Chart, bar chart

Description automatically generated

Figure 6:Assessment of Labour Leader Blair

Labour leader has got most frequent rating of 4 in the surveys.

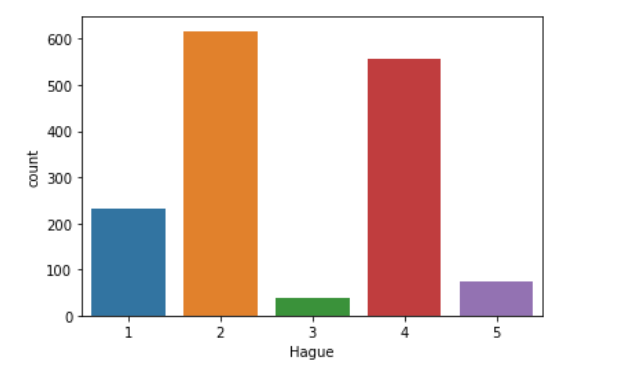


Figure 7:Assessment of Conservative Leader Hague

Conservative leader has got Most frequent rating of 2.

Chart, bar chart

Description automatically generated

Figure 8:Political Knowledge

Most frequent scale for political knowledge is 2.

Chart, bar chart

Description automatically generated

Figure 9:Gender

From the Graph there are 812 No. of females & 713 are males.

**Bivariate and Multivariate Analysis**

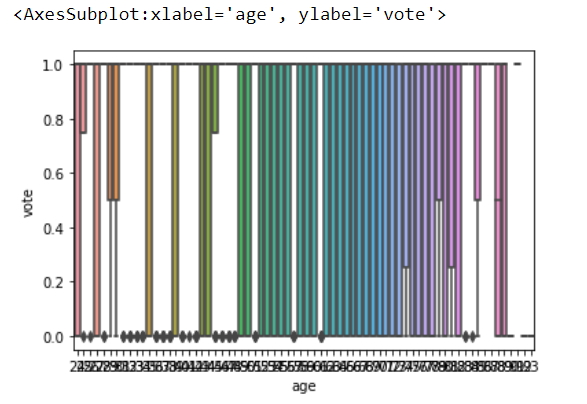


Figure 10:Age vs Vote

Above plot represents Labour Party is getting More Votes in each Age Group

Chart, bar chart

Description automatically generated

Figure 11:Current National economic conditions Party wise

Above graph shows that Labour party got better ratings.

Chart, bar chart

Description automatically generated

Figure 12:Current Household economic conditions party wise

As per above graph Conservative party have got less ratings in comparison to labour party

Chart, bar chart

Description automatically generated

Figure 13:Labour Leader Assessment

Labour leader has gotten maximum 4 ratings in their assessment.

Chart, bar chart

Description automatically generated

Figure 14:Conservative Leader Assessment

Conservative leader which are given good ratings as per the graph most of the time, but vote was given to Labour party leader

Chart, bar chart

Description automatically generated

Figure 15:Vote count in Europe

Above graph shows that most of the Voters have Eurosceptic attitude towards European integrations of Conservative Party

Chart, bar chart

Description automatically generated

Figure 16:Political Knowledge vs Vote

Labour part is having less Europeans integration so they have got most of the votes.

Chart, bar chart

Description automatically generated

Figure 17:Gender wise Voting

More of the males and females have given vote to the Labour party

Chart, scatter chart

Description automatically generated

Figure 18:Vote vs Age

In 20s There are very few peoples which have given Vote to the Conservative party.

Graphical user interface

Description automatically generated with low confidence

Figure 19:Vote vs Blair (Labour Leader)

Amongst labour party leaders which are given average ratings of 4 , Conservative party leaders have won there.

• Boxplot to check outliers –

Chart

Description automatically generated

Figure 20:Outlier Check

As mentioned earlier, we have found outliers in economic national conditions & household attributes, but upon checking further we understand that the ratings can be 1. Outliers are the datapoints which are extreme or captured mistakenly – so the data given is correct and it is not wise to treat these outliers.

Correlation:

Chart

Description automatically generated with medium confidence

Figure 21:Correlation

There is not a very strong Correlation amongst any of the variables of the dataset.

Pair plot:

Chart, box and whisker chart

Description automatically generated

Figure 22:Pair Plot

**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).**

As we know that we are having 2 Categorical Variables ‘Gender’ & ‘Vote’ so we have done one hot encoding with dropping of first column to avoid multicollinearity. As we know that Vote is Our Target Column, we have to classify whether a person have voted for Conservative or Labour party so for the sake of our better interpretation of model, we are not doing any encoding there.

Also we know that Age Group ranges from 20 to 100 and all other variables most of them are Ordinal Variables like rating ['vote', 'economic.cond.national', ' economic.cond.household,'Blair', 'Hague', 'Europe', 'political.knowledge', 'gender'], So for better understanding and interpretation of the Model we are Doing Binning of Age Column as below :

labels= ['20s', '30s', '40s','50s','60s','70s','80s','90s'])

Now we convert the above Values into category, and then we have performed Ordinal encoding for Changing it to numerical values for modelling in our Data.

Is Scaling necessary here or not?

As we can refer below Data range graph and can notice that data ranges are lying between 1 to 11 & most of them are ordinal so it has no meaning to scale the ordinal variables, So we are not doing scaling in this case.

Data Split: Split the data into train and test (70:30)

First, we have separated our target variable from the data, and we split the data in train test of (70:30) ratio by using python train\_test\_split function.

For every model we must first train that model and then test that model. Hence, we have split the data into train and test by passing below parameters into train test split function

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.30 , random\_state=0)

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

Logistic Regression We have applied logistic regression by passing following parameters

**• LR – Train data classification report**

**Table

Description automatically generated**

**• LR – Test data classification report**

**Table

Description automatically generated**

* **Confusion matrix for Test data in LR model**

****

* **Confusion Matrix for Training data in LR model**

**Chart

Description automatically generated**

**• Apply LDA (Linear Discriminant Analysis)**

We have applied LDA Function by passing below parameters –

LDA\_model= LinearDiscriminantAnalysis() LDA\_model.fit(X\_train, y\_train)

**• LDA Train Classification report**

**Table

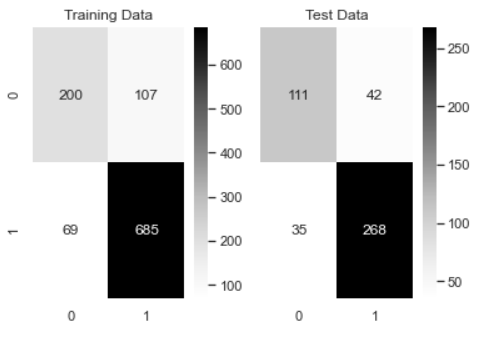
Description automatically generated**

**• LDA test Classification report**

**Table

Description automatically generated**

* **Confusion Martrix for train and test data in LDA Model**

****

Looking at recall & precision, accuracies are comparable in training and test data set for so we will fine tune our model by iterating various cut off probabilities and used gridsearchCV for fine-tuning the model.

**Inferences-LDA Model**

1.Order of importance of predictors-"Hague">"Blair">"political.knowledge">"economic.cond.national">"Europe".

2.Predictors "age","gender" and "economic.cond.household" are least important

3.Train accuracy and test accuracy & ROC-AUC scores are very similar ,hence this model do not suffers from overfitting .

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

**• KNN -Train Data classification report**

**Table

Description automatically generated**

**• KNN-Test Data Classification Report**

**Table

Description automatically generated**

**• Apply Naïve Bayes Model**

We have applied Naïve Bayes Model. by passing following parameters =-

NB\_model = GaussianNB() NB\_model.fit(X\_train, y\_train)

**• NB -Test Data classification report**

**Table

Description automatically generated**

**• NB -Training Data Classification Report**

**Table

Description automatically generated**

In the above Naïve Bayes model, we have used default parameters – They are giving comparatively low accuracy so we will change the value of default parameters to finetune the model.

**• Confusion Matrix of Naïve Bayes Model Training data**

**Text

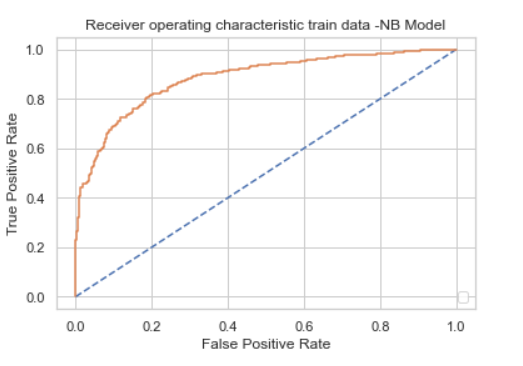
Description automatically generated with low confidence**

**• Confusion Matrix of Naïve Bayes Model Test data**

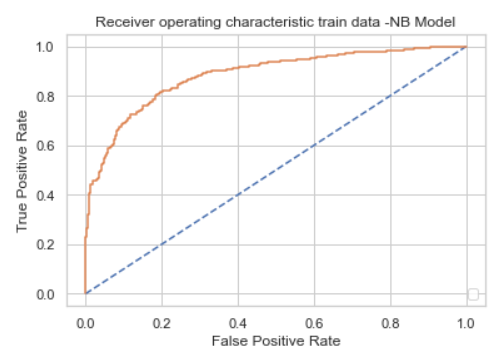
**Text

Description automatically generated with low confidence**

**• ROC Curve for Naïve Bayes Base Model Test data**

****

**• ROC Curve for Naïve Bayes Base Model Training data**

****

* **Comparison of KNN and Naïve Bayes Model Training data**

**Chart

Description automatically generated**

* **Comparison of KNN and Naïve Bayes Model Test data**

**Chart

Description automatically generated**

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

**• KNN Tuning**

We have performed KNN model in previous questions by choosing the best Kvalue where misclassification error should be minimal. So, we have calculated errors for various k values and plotted them as below –

Chart, line chart

Description automatically generated

**Formula for MCE Misclassification error**

(MCE) = 1 - Test accuracy score.

As we have seen from above plot misclassification Error is minimal on K=17 value. This means we will find our best accuracy by considering 17 nearest neighbours

**• Bagging (Random Forest should be applied for Bagging)**

**•Bagging RF- Train Data Classification Report**

**Table

Description automatically generated**

**•Bagging RF- Test Data Classification Report**

**Table

Description automatically generated**

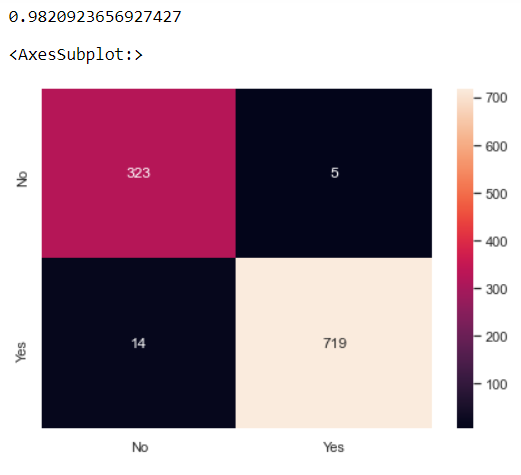
With random forest Bagging we found out that Train data Accuracy is 96% and test data accuracy is 83% so there will be overfitting. Because our model is performing better Train set but comparatively less performing in Test Data.

**• Boosting: -**

We have Used Ada-Boost and Gradient Boost for modelling , AdaBoostClassifier for Adaboost and Gradient Boost Classifier for Gradient boost.

**Before Gradient Boosting**

**Confusion Matrix for train set**

****

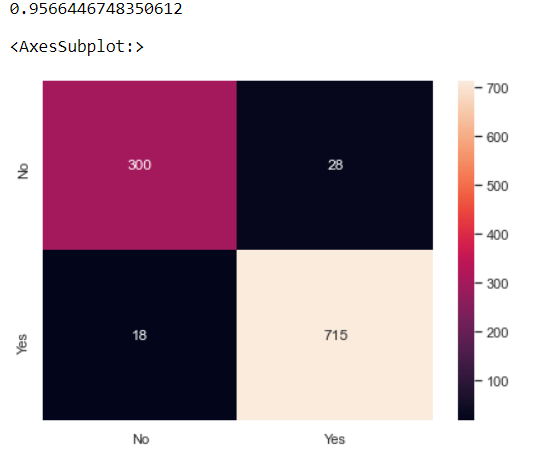
**Confusion Matrix for test set**

**Chart

Description automatically generated**

**After Gradient Boosting**

**Confusion matrix for Train set**

****

**Confusion matrix for Test set**

**Graphical user interface

Description automatically generated**

we found that Boosting & Bagging with Random Forest accuracy is higher. Also, we can make inference that bagging model is overfitted model and Gradient Boost is performing well in this case.

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

As per our business problem we have to predict which party a voter will vote for on the basis of the given information, to solve this problem we have made Logistic Regression , Linear Discriminant Analysis Model, K-Nearest neighbour & Naïve Bays Model to find out the best results LDA is performing slightly better than other models in terms of accuracy.

As we have seen from heat map that there is very less correlation between the variables which is good for the model. Referring to graphs in EDA we can notice the following –

• Voters between Age group of 30s to 70s are voting more.

• Voters in their 20s & 80s, 90 are voting significantly very less

• Significantly More no. of females have voted for Labour party

• Most of the Peoples, who have Eurosceptic attitudes given vote to the Labour Party.

• Labour party has got more votes than Conservative party.

**Recommendations** –

1. Collect more data like ratings on their previous leadership qualities (How they have performed previously), Religion of the respondent etc. to gain more insight.

2. CNBE can take Online surveys so that it can reduce their actual cost on surveys in res ult they can collect more data.

3. CNBE can also give free eBooks or online Coupons to the voters if they participate in surveys.

4. Company can collect the ratings on the attitude of leader towards Current issues.

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.

We have imported the three text files from inaugural fieldIDs.

1. inaugural.raw('1941-Roosevelt.txt')

2. inaugural.raw('1961-Kennedy.txt')

3. inaugural.raw('1973-Nixon.txt')

We are assigned these text files into variable called ‘x’ converting them to list files. After that we have to call these text files into a new dataframe. So that we can easily perform the rest of the task. Let us call the new dataframe as ‘y’. The speech inside each files are converted into the list format and applied to the text column of the ‘y’ dataFrame.

**Graphical user interface

Description automatically generated with medium confidence**

2.2 Remove all the stopwords from all three speeches.

For Roosevelt speech we have around 632 stopwords.

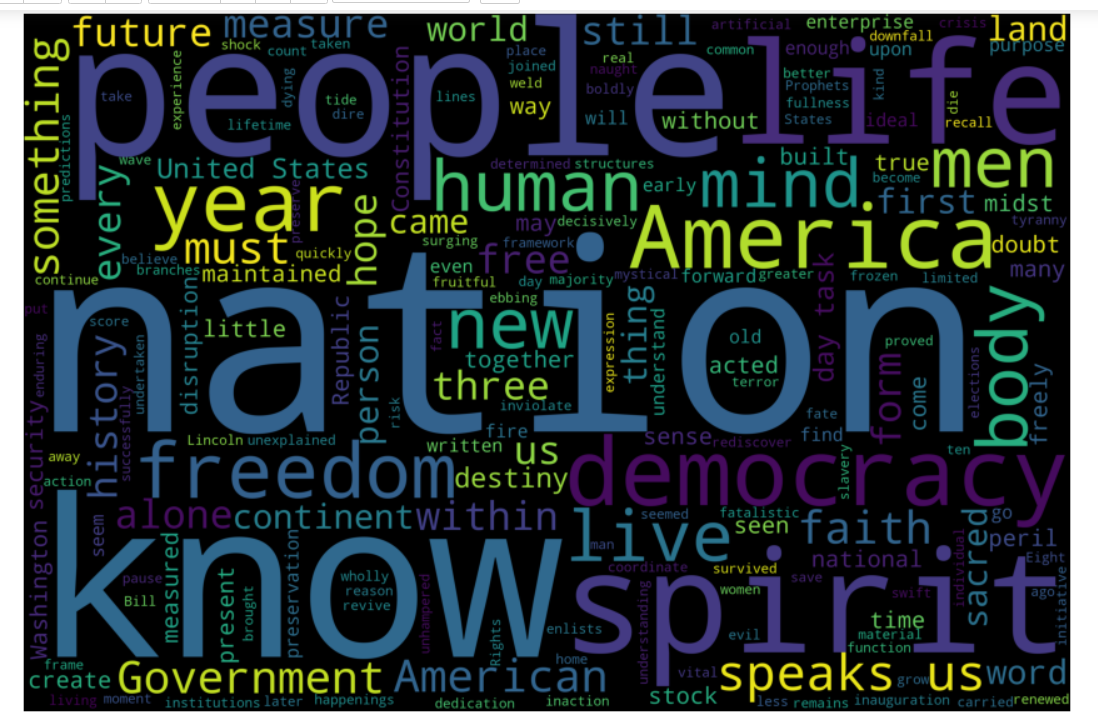
We have to remove all the stopwords and punctuation from the speech to find out the frequent words in the speech.

From nixon speech we got around 899 stopwords. And for kennedy speech there are around 618 stopwords.

Before removing the stopwords, we are first converting the speech files into lower cases and splitting them into words. After those special characters, numerical values and punctuations are removed from the text files, using nltk corpus library we are importing stopword function and calling them to remove stopwords.

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

**Roosevelt**

****

**Kennedy**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Nixon**

**Text

Description automatically generated**