**MACHINE LEARNING PROJECT**

**Gaurav Sood**

**PGP-DSBA Online Dec\_B’21**

**Table of Contents**

**Problem Statement……………………..……………………………………………………………….5**

1. Reading the Dataset 6

2. EDA, Univariate & Bivariate Analysis 10

3. Data Encoding 21

4. Logistic Regression & LDA 23

5. KNN & Naïve Bayes Model 28

6. Model Tuning, Bagging & Boosting 36

7. Performance Metrics 42

8. Inference 43

**List of Tables**

Table no.1: Head of the Dataset .……....…………………………………………………………………………….…………………. 06

Table no.2: Tail of the Dataset ……………………………………………………………………………………………………….……. 06

Table no.3: Data Info …………………………………………………………………………………………………………………………… 07

Table no.4: Dataset Sample ……………………………….………………………………………………………………………………… 07

Table no.5: Null Value Check ……..………………………………………………………………………………………………………… 08

Table no.6: Data Description ..……………………………………………………………………………………………………………… 08

Table no.7: Vote Value Count ……………………….…….………………………………………………………………………………. 10

Table no.8: Gender Value Count …………….…………………………………………………………………………………………… 10

Table no.9: Dummy Variables ………….………………………………………………………………………………………………..… 22

Table no.10: Transformed Variables .…………………………………………………………………………………………………… 22

**List of Figures**

Figure no. 1: Vote Graph ..………………………………………….……………………………………………………………………….. 11

Figure no. 2: Gender Graph ……………...…………………………………………………………………………….…………………… 11

Figure no. 3: Age Distribution Graph …..……………………………………………………………………………..……………….. 12

Figure no. 4: Blair Graph …………………….……………………………………………………………………………………………….. 12

Figure no. 5: Hague Graph …………………….…………………………………………………………………………………………….. 13

Figure no. 6: Europe Graph …………………………………………………………………………………………………………………. 13

Figure no. 7: Political Knowledge Graph ……………..………………………..…………………………………………………….. 14

Figure no. 8: Household Economics Graph ……………………….………..……………………………………………………….. 15

Figure no. 9: National Economics Graph ……………………………………………………………………………………………… 15

Figure no. 10: Vote vs. Blair Graph ….………………………………………………………..………………………………………… 16

Figure no. 11: Vote vs. Hague Graph …………………………………………..…….………………………………………………… 16

Figure no. 12: Gender vs. Vote Graph ..........………………………………………………………………………………………… 17

Figure no. 13: Vote vs. Europe Graph ………………………………..………………………………………………………………… 18

Figure no. 14: Vote vs. Political Knowledge Graph ……………………………………………….……………………………… 18

Figure no. 15: Correlation Heat-map …………..……………………………….……………………………………………………… 19

Figure no. 16: Outlier Boxplot ……………………………………………………………………………………………………………… 20

**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.

**Data Dictionary**

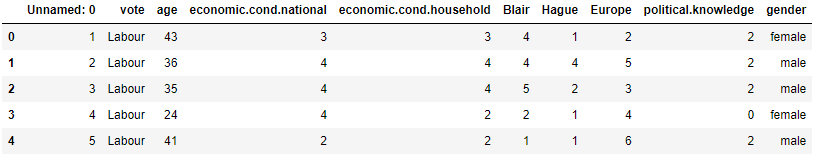
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1. ***Vote***: Party choice: Conservative or Labour | | | | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. ***Age***: in years | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. ***Economic.cond.national*:** Assessment of current national economic conditions, 1 to 5. | | | | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. ***Economic.cond.household*:** Assessment of current household economic conditions, 1 to 5. | | | | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. ***Blair***: Assessment of the Labour leader, 1 to 5. | | | | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6. ***Hague***: Assessment of the Conservative leader, 1 to 5. | | | | | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7. ***Europe***: an 11-point scale that measures respondents' attitudes toward European integration. High scores  represent ‘Eurosceptic’ sentiment. | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8. ***Political.knowledge*:** Knowledge of parties' positions on European integration, 0 to 3. | | | | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 9. ***Gender***: female or male. | | |  |  |  |  |  |  |  |  |  |  |  |

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

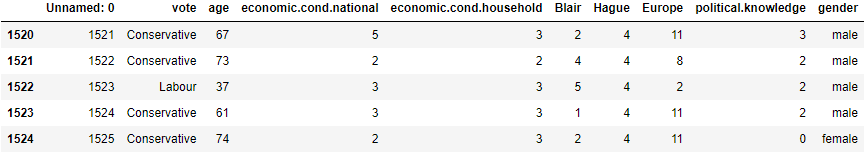
First, we load all the necessary libraries for model building. Then, we import the dataset provided.

Following this, we read the head and tail of the dataset to check whether the data has been properly fed.

**Sample of the Dataset**



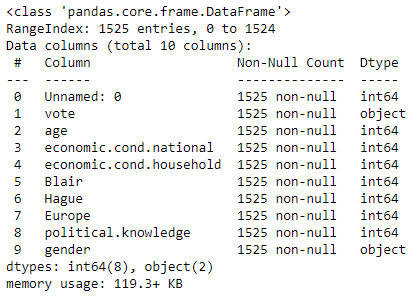
**Table no. 1: Head of the Dataset**



**Table no. 2: Tail of the Dataset**

As we can see, the data has been imputed properly.

**Exploratory Data Analysis**

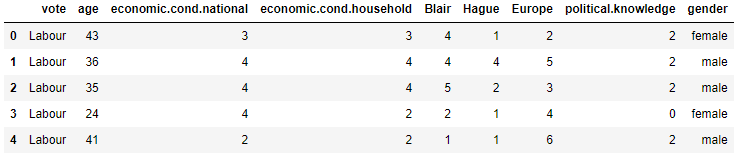


**Table no. 3: Data Info**

As we can see from the above table, the given dataset has 1525 entries across 10 columns. All the variables are of integer type, except ‘vote’ and ‘gender’ which are of object type.

We will now drop the unnamed column, as it is not relevant as we proceed further.

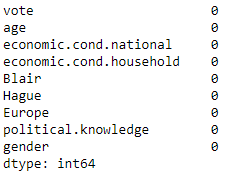
**Checking the Dataset**



**Table no. 4: Dataset Sample**

The column has been successfully dropped.

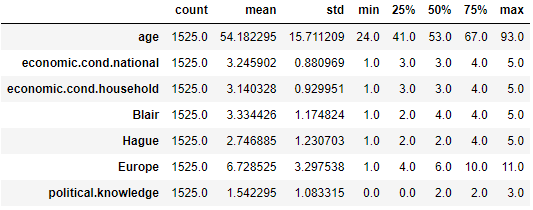
**Checking for Null Values**



**Table no. 5: Null Value Check**

There are no null values in the given dataset.

**Data Description**



**Table no. 6: Data Description**

A number of observations can be made from the above descriptive statistics:

* From the voter turnout, the youngest voter is 24 years of age while the oldest is aged 93.
* The average voter in the survey is about 54 years old.
* Assessment of the current national economic condition seems to be about average (approximately 3 on a scale of 1-5).
* Assessment of the current household economic conditions amongst the voters also seems to be average (approximately 3 on a scale of 1-5).
* The Labour party leader ‘Blair’ is generally more favoured than the Conservative party leader ‘Hague’.
* A general consensus could not be found as it pertains to the ‘Europe’ issue. The voters surveyed cannot be said to be leaning towards being ‘Eurosceptic’ or ‘Eurocentric’.
* The voters’ political knowledge when it comes to their party’s stance on European integration was found to be about average.

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

We made some key findings from the descriptive statistics check above. However, we can also understand a lot by looking at the object variables ‘vote’ and ‘gender’ individually.

**Vote Value Count**



**Table no. 7: Vote Value Count**

The Labour voters are outnumbering the Conservative voters by a large margin in the given surveyed dataset (more than two times over).

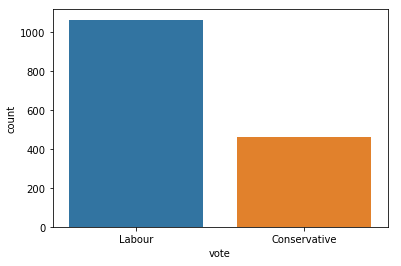
**Gender Value Count**



**Table no. 8: Gender Value Count**

There are more female voters than male voters in the dataset. This could possibly mean that women’s issues are being represented well by either or both of the parties in the election.

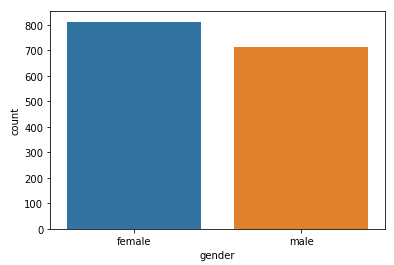
**Vote Plot**



**Figure no.1: Vote Graph**

As mentioned above, there are much more Labour party voters than Conservative party voters in the given dataset.

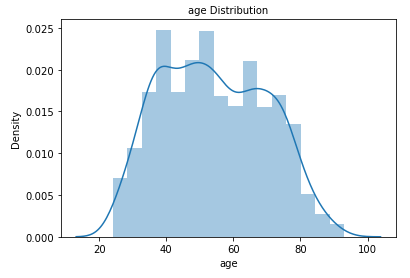
**Gender Plot**



**Figure no.2: Gender Graph**

As mentioned above, there are slightly more female voters than male voters in the given dataset.

**Age Distribution Plot**



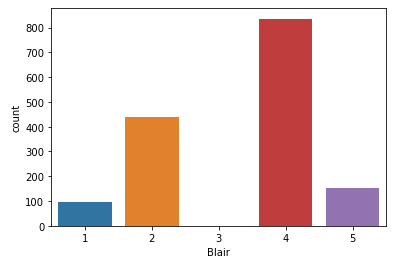
**Figure no.3: Age Distribution Graph**

There seems to be a large turnout amongst middle-aged voters according to the surveyed data.

Senior citizens have also showed up in numbers to vote in the election.

Conversely, not many young voters are observed in the dataset.

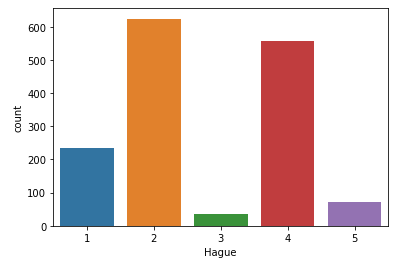
**Blair Plot**



**Figure no.4: Blair Graph**

The Labour party candidate ‘Blair’ seems to have a high approval rating amongst the surveyed voters, with over 800 of the 1500 polled giving him a 4 out of 5 grade. There were over 400 people than graded him a 2 out 5 however, and none graded him a median score of 3 out of 5, indicating that there are not many swing voters in the dataset.

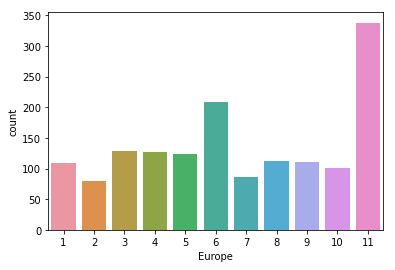
**Hague Plot**



**Figure no.5: Hague Graph**

The Conservative party candidate ‘Hague’ has mostly negative approval ratings, with over 600 voters grading him 2 out of 5 and over 200 going as far as giving him a 1 out of 5. There seem to be some loyal Conservative voters however, as over 500 voters did grade him 4 out of 5.

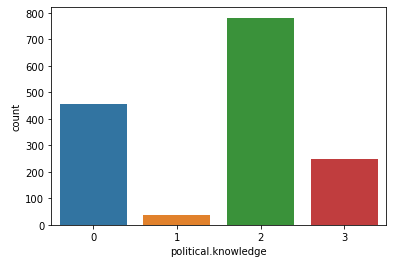
**Europe Plot**



**Figure no.6: Europe Graph**

A large number of voters polled are found to be highly Eurosceptic (over 300 of them). Around 200 seem to be on the fence on the Europe issue. The rest were found to be evenly distributed amongst our 11-point scale.

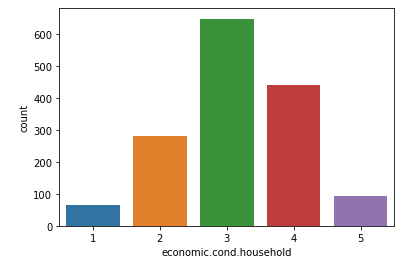
**Political Knowledge Plot**



**Figure no.7: Political Knowledge Graph**

Most voters were found to be well versed in their understanding of their respective parties’ position on European integration. However, a few (over 400) did not seem to have any idea on where their party stands when it comes to the Europe issue.

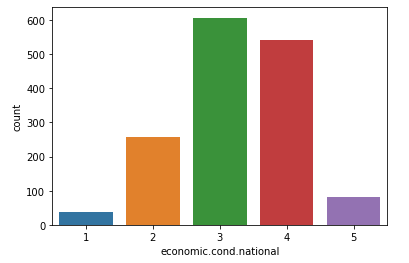
**Household Economic Condition Plot**



**Figure no.8: Household Economics Graph**

The bulk of the voter base was found to be doing just fine economically, with most getting an average household economic condition grade of 3 out of 5. There were a few on the extremes however, with about a 100 each doing very poorly (1 out of 5) or very well (5 out of 5).

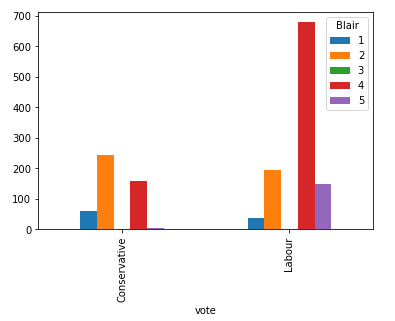
**National Economic Condition Plot**



**Figure no.9: National Economics Graph**

Most voters polled were of the idea that the national economy is doing quite well when it comes to the economic side of things. This graph almost mirrors the above household economics graph, but we cannot say that the same voters are giving similar responses to both the household economics as well as the national economics question.

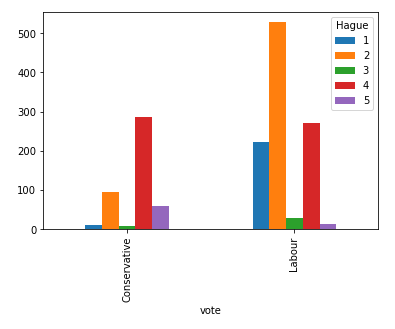
**Vote vs. Blair Plot**



**Figure no.10: Vote vs. Blair Graph**

As expected, Labour voters have a favourable rating of Blair and Conservative voters have a generally unfavourable rating of Blair.

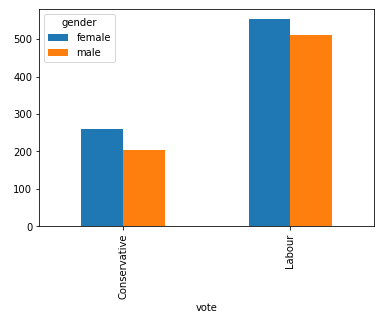
**Vote vs. Hague Plot**



**Figure no.11: Vote vs. Hague Graph**

As expected, Conservative voters have a favourable rating of Hague and Labour voters have an unfavourable rating of Hague.

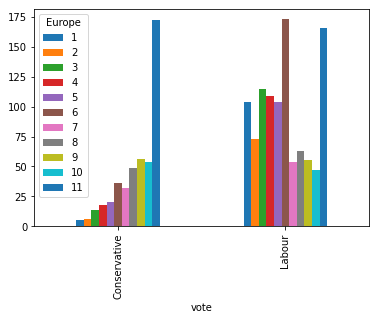
**Relationship between Gender and Vote**



**Figure no.12: Gender vs. Vote Graph**

There does not seem to be any relationship between the gender of the voter and the party they vote for.

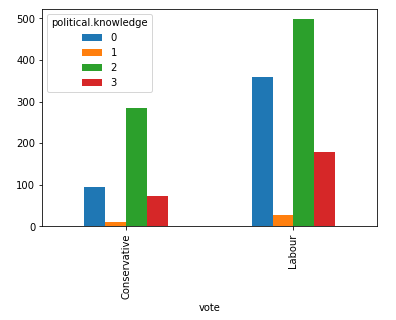
**Relationship between Vote and Europe**



**Figure no.13: Vote vs. Europe Graph**

Conservative voters were found to be leaning towards being Eurosceptic whereas their Labour counterparts’ opinions were mostly all over the place.

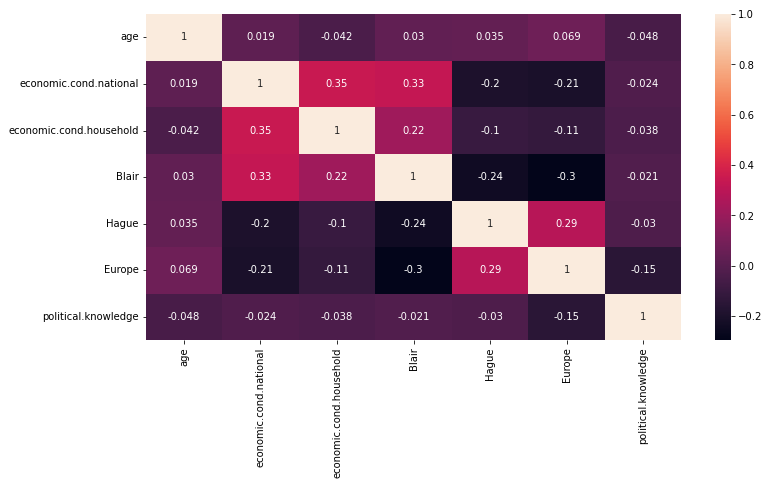
**Relationship between Vote and Political Knowledge**



**Figure no.14: Vote vs. Political Knowledge Graph**

These variables are in proportion for the most part, however, Labour voters are more unaware of their parties’ stance on European integration than their Conservative counterparts.

**Correlation Heat-map**

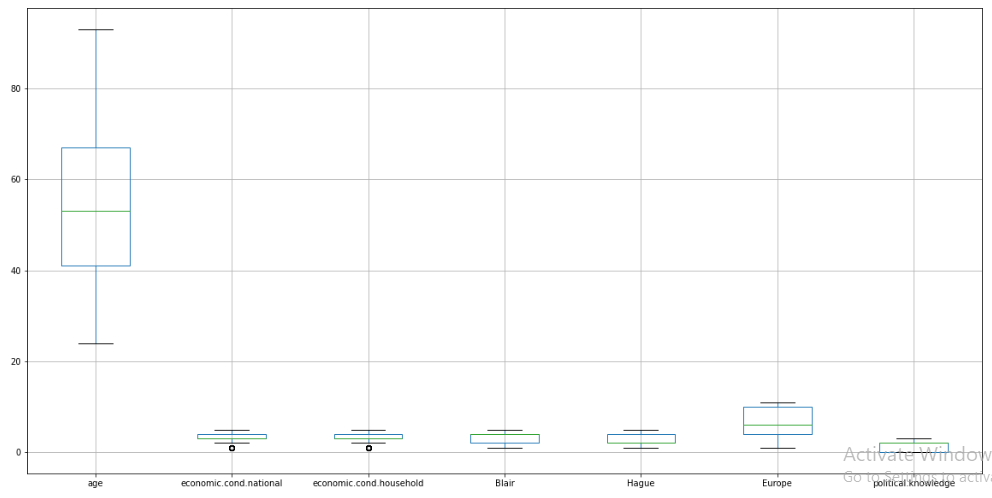


**Figure no.15: Correlation Heat-map**

Some key observations can be made from the above heat-map:

* There is some correlation between those having a favourable opinion of Hague vs. those grading high on the 11-point Europe scale. This verifies our earlier claim that Conservative voters are more Eurosceptic than their Labour counterparts.
* There is also some correlation between the variables ‘Economic condition household’ and ‘Economic condition national’. This indicates that those people who are doing well financially think that the nation as a whole is heading in the right direction when it comes to economic matters.
* Another interesting observation is that there is some relationship between ‘Blair’ and ‘Economic condition national’ (and subsequently ‘Economic condition household’). This could mean that those that are in agreement with the country’s financial policy and think they are progressing economically generally favour Blair. Also, we can make an assumption that upper class and middle class people are supporting Blair, whereas blue-collar workers are generally Conservative and voting for Hague.

**Outlier Boxplot**



**Figure no.16: Outlier Boxplot**

There are little to no outliers present in the given dataset.

The only outliers found were in the variables ‘Economic Condition Household’ and ‘Economic Condition National’. These are the ones who belong to the poorest section of the society, and those who think that the nation is in (or heading towards) financial ruin, respectively.

1. **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).**

There is a crucial step to be done before the modeling process and that is to check the data for missing values. Missing values can bias the results of the machine learning models and/or reduce the accuracy of the model.

We have done so earlier and found no such missing values. Hence, we can begin with the data encoding process.

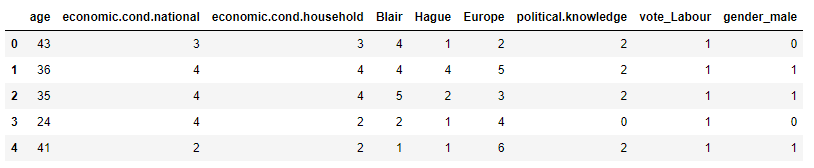
We divided the variables based on their typing:

1. Continuous variables, which are of float and integer type; and
2. Categorical variables, which are of object type.

**Creating Dummy Variables**

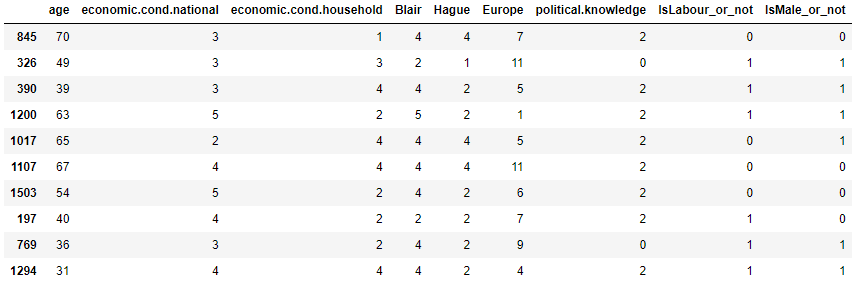
After checking the number of categorical values in each column, we move on to create dummy variables in binary or in multi classification variables for each categorical variable to help us to predict better.

This step is crucial, as categorical variable transformation is important for any models and its selection matters a lot for model performance. It is a way to make the data work better in our model.



**Table no. 9: Dummy Variables**

Here, we have created dummy variables in binary form for each of the columns ‘Vote’ and ‘Gender’, where 1 corresponds to a Labour vote and Male gender for each of the columns, respectively. We shall next rename these columns.



**Table no. 10: Transformed Variables**

Here, we have transformed 2 variables, ‘Vote’ and ‘Gender’ to work in our Model.

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

Here we are splitting the data into training and testing model, where training data is 70% and testing data is 30%.

**Linear Discriminant Analysis**

**Train Set**

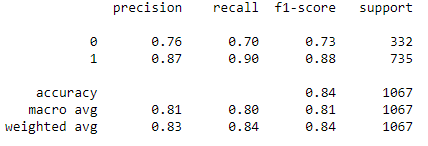
LDA Model Score: 0.8369259606373008

**Confusion Matrix**

[[233 99]

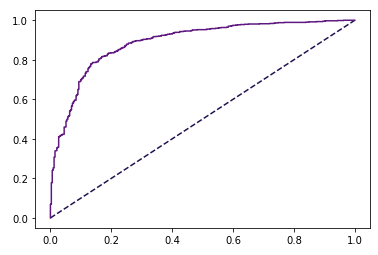
[ 75 660]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.889



**Test Set**

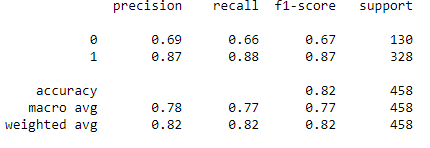
LDA Model Score: 0.8187772925764192

**Confusion Matrix**

[[ 86 44]

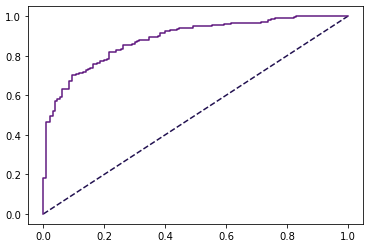
[ 39 289]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.884



**Logistic Regression Model**

**Train Set**

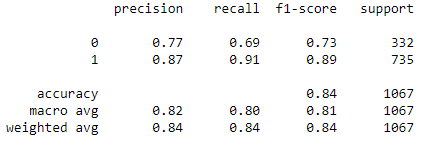
Logistic Model Score: 0.8406747891283973

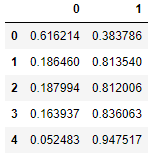
**Confusion Matrix**

[[230 102]

[ 68 667]]

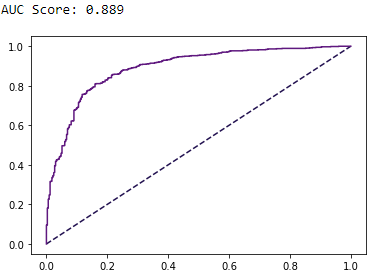
**Prediction Table**





Logistic Model Score: 0.8406747891283973

**ROC Curve**



**Test Set**

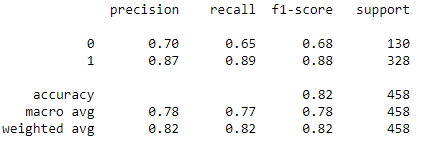
Logistic Model Score: 0.8231441048034934

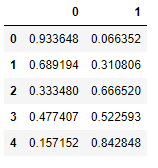
**Confusion Matrix**

[[ 85 45]

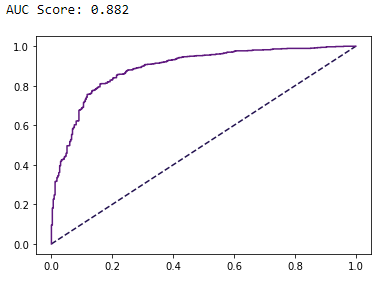
[ 36 292]]

**Prediction Table**





**ROC Curve**



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

**KNN Model**

**Train Set**

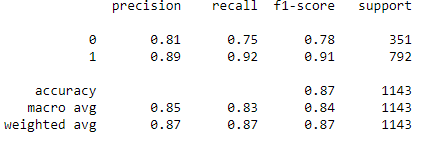
KNN Model Score: 0.8678915135608049

**Confusion Matrix**

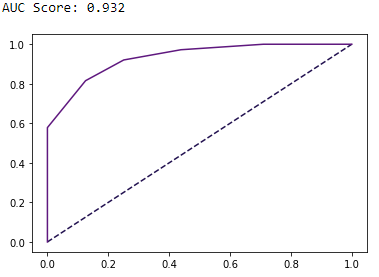
[[263 88]

[ 63 729]]

**Prediction Table**



**ROC Curve**



**Test Set**

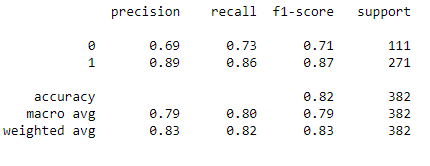
KNN Model Score: 0.824607329842932

**Confusion Matrix**

[[ 81 30]

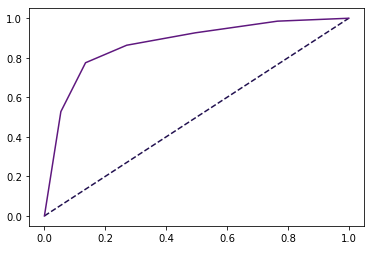
[ 37 234]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.870



**KNN Model (For no. of neighbours = 7)**

**Train Set**

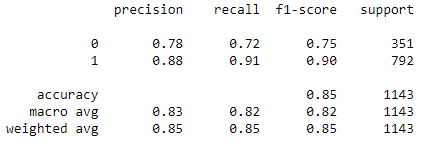
KNN Model Score: 0.8530183727034121

**Confusion Matrix**

[[253 98]

[ 70 722]]

**Prediction Table**



**Test Set**

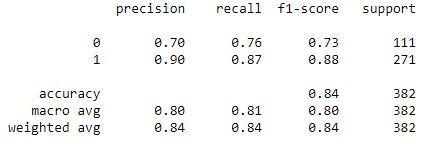
KNN Model Score: 0.8350785340314136

**Confusion Matrix**

[[ 84 27]

[ 36 235]]

**Prediction Table**



**KNN Model (For no. of neighbours = 5)**

**Train Set**

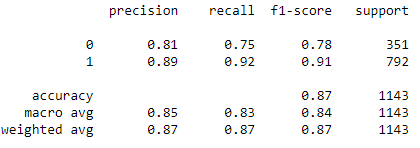
KNN Model Score: 0.8678915135608049

**Confusion Matrix**

[[263 88]

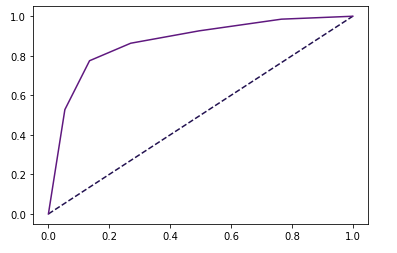
[ 63 729]]

**Prediction Table**



**ROC Curve**

AUC Score 0.904



**Test Set**

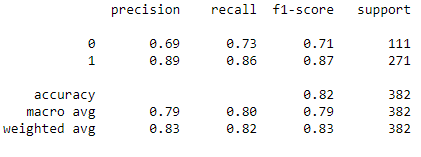
KNN Model Score: 0.824607329842932

**Confusion Matrix**

[[ 81 30]

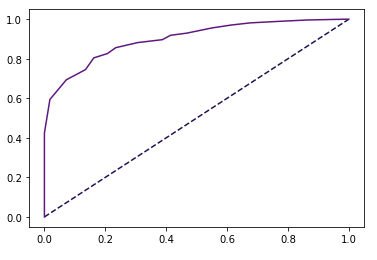
[ 37 234]]

**Prediction Table**

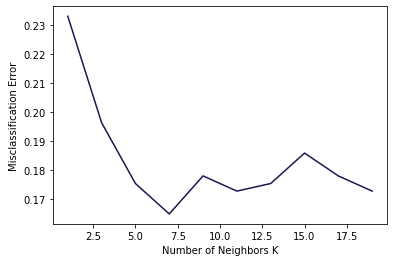


**ROC Curve**

AUC Score: 0.900



**No. of Neighbours vs. Misclassification Error**



**Naïve Bayes Model**

**Train Set**

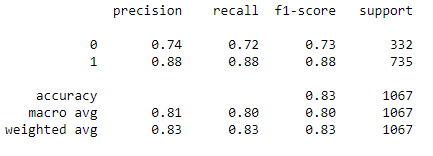
NB Model Score: 0.8331771321462043

**Confusion Matrix**

[[240 92]

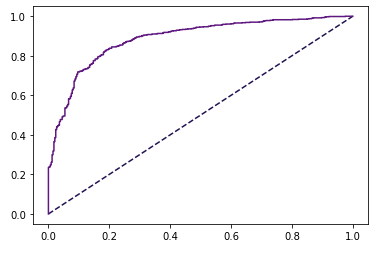
[ 86 649]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.886



**Test Set**

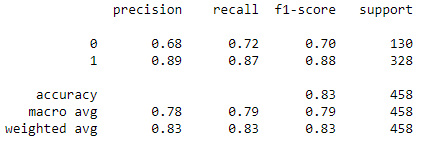
NB Model Score: 0.8253275109170306

**Confusion Matrix**

[[ 94 36]

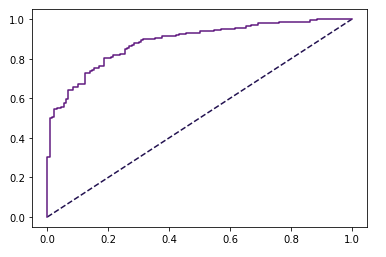
[ 44 284]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.885



1. **Model Tuning, Bagging and Boosting.**

**Bagging**

**Train Set**

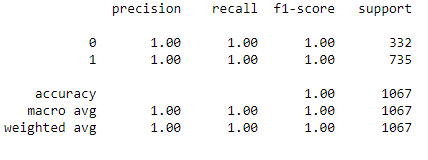
Bagging Model Score: 0.9990627928772259

**Confusion Matrix**

[[331 1]

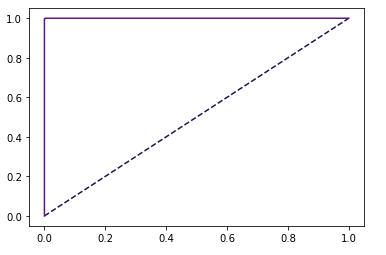
[ 0 735]]

**Prediction Table**



**ROC Curve**

AUC Score: 1.000



**Test Set**

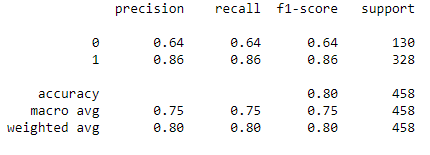
Bagging Model Score: 0.7969432314410481

**Confusion Matrix**

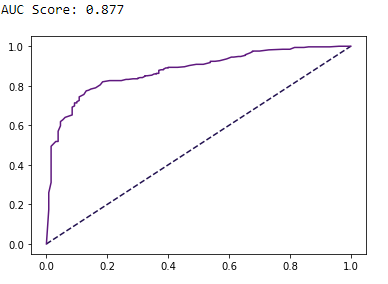
[[ 83 47]

[ 46 282]]

**Prediction Table**



**ROC Curve**



**ADA Boosting vs. Gradient Boosting**

**Train Set**

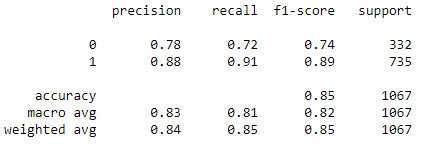
ADB Model Score: 0.8472352389878163

**Confusion Matrix**

[[238 94]

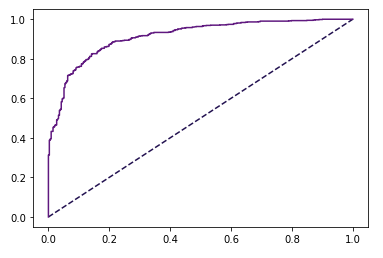
[ 69 666]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.913



**Train Set**

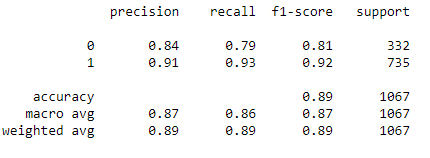
GBC Model Score: 0.8865979381443299

**Confusion Matrix**

[[240 92]

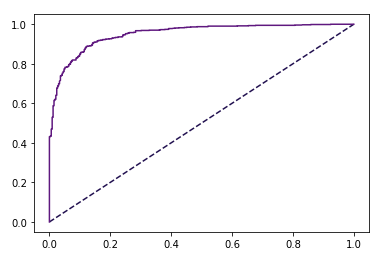
[ 86 649]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.950



**Test Set**

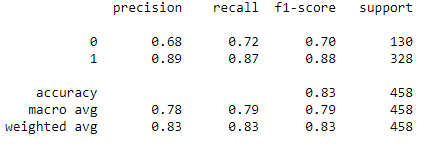
ADB Model Score: 0.8187772925764192

**Confusion Matrix**

[[ 94 36]

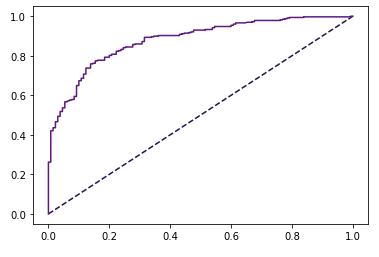
[ 44 284]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.879



**Test Set**

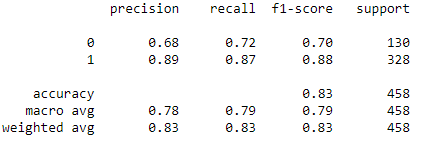
GBC Model Score: 0.8318777292576419

**Confusion Matrix**

[[ 94 36]

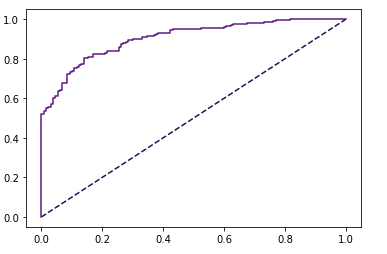
[ 44 284]]

**Prediction Table**



**ROC Curve**

AUC Score: 0.904



1. **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.**

We have checked the performance metrics above for each and every model using all the available performance metrics (Accuracy, Confusion Matrix, ROC Curve plots and obtaining AUC Scores).

We can select any model from the above few models that we have worked on. However, the Bagging model gave us a perfect 100% Accuracy on its Training Set. This could be due to overfitting and result in poor generalization. So, we will disregard the Bagging model.

We will go ahead and select the Gradient Boosting Model, as it gives us great scores throughout all the metrics, with Accuracy scores of 88.6% and 83.1% on its Train and Test sets respectively, and AUC Scores of 95% and 90.4% respectively.

We could have selected another among the many models that we built, such as Naïve Bayes Model, but Gradient Boosting Classifier is highly accurate, has faster training speed and is more efficient. It excels at classification and regression problems, hence is more practical.

1. **Based on these predictions, what are the insights?**

After performing extensive EDA and building different models and comparing/contrasting them, we have gained a number of actionable insights:

* The Gradient Boosting Model was chosen as the desired and preferred model through our analysis. This Model offered us excellent scores in each of the performance metrics (Accuracy, AUC Score, AUC\_ROC Curve, Recall Value, etc.)
* The Labour party is outperforming the Conservative party by a huge margin. They have two and a half times the votes of the Conservative party.
* Female voter turnout is greater than that of male voters. One of the reasons for this could be women’s issues being well represented by either or both of the parties.
* The Conservative party is filled with people having higher Eurosceptic sentiments. The Labour party has a good mix of Eurocentric voters and Eurosceptic voters.
* Conservative voters are found to be more political aware in terms of their political knowledge than their Labour counterparts.
* Those who are financially settled and well-off are inclined to vote for Blair, whereas blue-collar voters lean towards Hague.

**THE END**