**DALARNA UNIVERSITY 2020**



**STATISTICAL LEARNING**

**HOME EXERCISE - 1**

**KIRAN SAI TANNEERU**

## **h19kirta@du.se**

**Introduction:**

A data set has been given to us regarding a survey data collected during the 2016 US prudential election campaign. For the tasks based on this data set we have to perform different analysis and produce our answers with reasoning. The tasks are as follows:

1. Recode the variable Trump as follows. Denote Slightly liberal to Extremely liberal (levels 1-3) as “Liberal”, and Moderate to Extremely conservative (levels 4-7) as “Conservative”. Is there any personal characteristics of the individuals that determines whether someone would consider Donald Trump as Liberal (or conservative)? Motivate your methods and interpret your results.

Explanation:

The data set consists of 4271 entries and 18 columns representing various data fields of Trump and Hilary.

In this first task, we have been asked to assign levels 1 – 3 as liberal to extremely liberal and levels 4 – 7 as moderate to extremely conservative. Also, we need to determine whether trump was liberal or conservative with inclusion of any personal characteristics.

For this, I have created two sets liberal and conservative by dividing the whole data set into two subsets. Now I would be eliminating the null values and other values which are not required. After this we can check the data and we find the useful data only.

From the above plot, we can check the correlation between the variables.

A screenshot of a cell phone

Description automatically generated

From the above heat map we can find correlation between partner and spouseedu to be high, so we are going to remove partner variable from the model. Even though the heat map did not show any relation between trump and rest of variables it is still useful for variable selection in our model

Now we check the significance of the variables and now we can predict the probability of trump being liberal or conservative. Now we check the characteristics of trump to determine the character.

A close up of text on a white background

Description automatically generated

The above summary shows the significance of the variables regarding to Trump variable.

A screenshot of a cell phone

Description automatically generated

The above summary shows the association between the variable trump and rest of the variables. By looking at the significance level of the variables we can conclude how strong the association is between the response and the predictors. Now we produce a confusion matrix and get the correctly predicted values and we get 3151 times to be conservative against 172 times being liberal and we get an accuracy of 82.23%.

A close up of a person

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The above summary is done with only the highly significant variables. Now we produce a confusion matrix and get the correctly predicted values and we get 3147 times to be conservative against 174 times being liberal and we get an accuracy of 82.18%.

So, we can conclude that the personal characteristics has an effect on the variable trump being conservative or liberal.

2. Build a suitable prediction model to predict an individual’s party identification using the respective individual’s other personal, and family characteristics. Experiment with different methods, and model specifications, and motivate your choice.

Explanation:

For this task I have used two classification algorithms.

1. LDA
2. QDA
3. KNN

And to perform these, I have split my data into training and testing with 70%(2828) in training set and 30% in testing.

LDA: For LDA I am getting an accuracy of 48.3% with all variables into consideration and an accuracy of 44.6% from the subset model.

A picture containing many, different, large, people

Description automatically generatedA picture containing clock, room, orange, different

Description automatically generated

Above diagrams gives the confusion matrix of LDA in both cases.

QDA: For QDA I am getting an accuracy of 42.7% with all variables into consideration and an accuracy of 44.7% from the subset model.

A picture containing orange, room, clock, hanging

Description automatically generatedA picture containing room, clock, wooden, people

Description automatically generated

Above diagrams gives the confusion matrix of QDA in both cases.

KNN: For KNN I am getting a accuracy of 35.6% with a cluster size of 3, an accuracy of 37.5% with a cluster size of 8 and an accuracy of 37.1% with a cluster size of 10.

A screenshot of a cell phone

Description automatically generatedA picture containing bird, people

Description automatically generated

A close up of a black background

Description automatically generated

Above diagrams gives the confusion matrix of QDA in both cases.

According to the observations from LDA, QDA and KNN, we can say that LDA has higher accuracy in both the cases.