Introduction:

We have been given the survey data collected from the US Prudential Election campaign for the year 2016.

By following the above data set we need to solve the given specific task by using suitable methods for the data. The tasks are:

1. Recode the variable Trump from the data set as follows. Levels 1-3 as “Liberal” and Levels 4-7 as “Conservative”. And need to analyse that are there any personal characteristics of the individual that determine whether someone would be consider Donald Trump as Liberal or Conservative.

2. Build a respective prediction model to predict an individual party identification using the respective individual other personal and family characteristics.

Overview of the Data:

The dataset has 18 different variables which are been divided as the columns of the data and they are numerical encoded like if we considered the variable Partner in the dataset it has 4 different values which are -9,-1,1,2 each values would be representing different relations on their marital status.

Initial Steps:

Firstly, I have loaded the CSV file “ANES2016.csv” to the R Studio as a dataframe. The data set has 4271 entries as the rows for 18 columns which would be describing the personal characteristics of the individuals like Age, Education, Income, Family Size, and it also has the PartyID, Trump and Hilary. And all the variables in the dataframe are numerical coded in the form categorical variables.

Methods:

Task 1: We have been asked to recode the variables in the Trump column levels as 1-3 as “Liberal” and levels 4-7 as “Conservative”. Now, after the recode to have the specific dataframe, I have created a subset for the categorical variable “Liberal” and “Conservative” which would be eliminating the other values which are not required for the analysis. I would be also cleaning our data to eliminate the null values in the data frame which is required to perform an accurate analysis. Now the summary() function would be giving an overall impression of the numerical summary for the each variable in the data and we can also see that there are no NA values in the data frame as we have omitted the Null values during the cleaning stage of the data.

Method 1: Logistic Regression Model

Logistic regression models are a great tool for analysing binary and categorical data allowing you to perform a contextual analysis to understand the relationships between the variables, test for the difference, estimate effects, make predictions, and plan for future scenarios [1].

• Using the logistic regression model (glm), we would be finding the significance of the variables by calling the Trump variable as the Response variable. And I would be eliminating the column ID which is irrelevant column to check for the significance.

• Now, using the summary() function to verify all the aspects for the fitted model, which are p-values for the coefficients. The smallest the p-value the higher the significance between the coefficients and the response variable. Now, if we check the p-values for the fit, we see that Hilary, Age, SpouseEdu, GBirth, Dependent, Income, Education2 have some relatable significance on the response variable Trump.

• Now, to predict the probability that the Trump would be “Liberal” or “Conservative” we can use the predict function.

• Now, we need to convert the predicted models into class labels as Liberal or Conservative to make the necessary predictions whether Trump will be Liberal or Conservative on the characteristics of the person.

• Now, we can use the xtabs() function to produce the confusion matrix to determine the correct predicted values. The diagonal elements in the confusion matrix the correct predicted value and the off-diagonal values in the confusion matrix are the wrongly predicted values.

• If we check the confusion matrix, we can say we have correctly predicted Conservative for “3151” time and Liberal for “172” time.

• Now we can use mean() function to see the accuracy for the prediction. In this case, we can see that the correctly predicted accuracy is 82%.