Statistical Learning
Report
On
Home Assignment 1

HÖGSKOLAN
DALARNA

By
Manthri Bala Kiran
H19baman@du.se

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 data frame, 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: Finding Correlation

To find the correlation we need to use the data which has not been recoded as the Liberal and Conservative. Firstly, we can see that correlation by using the chat correlation for the columns. We can see that they are few variables which are highly correlated with each other like Partner & SpouseEdu and Employment & Dependent and Partner & Martial.
From the above correlation we are not able to find the relation for Trump with other variables. To find the correlation for the variable Trump we can use Classification methods like Logistic regression.

Method 2: 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%.

Task 2:

To classify the response variable which has more classes such as Logistic regression, Quadratic regression, and Linear Discriminant Analysis is useful where the classes are well separated, and the parameter distribution is stable.

Method 1:

- Now, I am going to fit an LDA model using the lda() function while considering PartyID as the response variable and all the other variables as the input variables.
- From the Prior probabilities we can see that 34% have been opted for Democrat, 28% opted for Republican, 32% opted for Independent, and while only 4% opted for other party.
- We can also see the group means with respective of the PartyID.
- The coefficients of linear discriminants provide the linear combination of input variables that are used to form the LDA decision rule.
- Now we use predict() function on the test data which will be returning three elements.
- Now we can use the confusion matrix to evaluate the performance of the model. We can see that 56% it has predicted correctly for Democrat, 62% correctly predicted for Republican, 29% correctly predicted for Independent.

- We can calculate the mean of the model by using the mean() function and it is noticed that it has correctly predicted for 48%.

Method 2:
QDA works as a quadratic decision boundary, it can accurately model a wider range of problems.

- Now, I would be fitting QDA by using the qda() function while considering PartyID as the response variable and the other variables as the input variable.
- From the Prior probabilities we can see that 34% have been opted for Democrat, 28% opted for Republican, 32% opted for Independent, and while only 4% opted for other party.
- We can also see the group means with respective of the PartyID.
- Now the predict() function to classify the test data works similar as for LDA
- Now we can use the confusion matrix to evaluate the performance of the model.
- We can calculate the mean of the model by using the mean() function and its is noticed that it has correctly predicted for 36%.

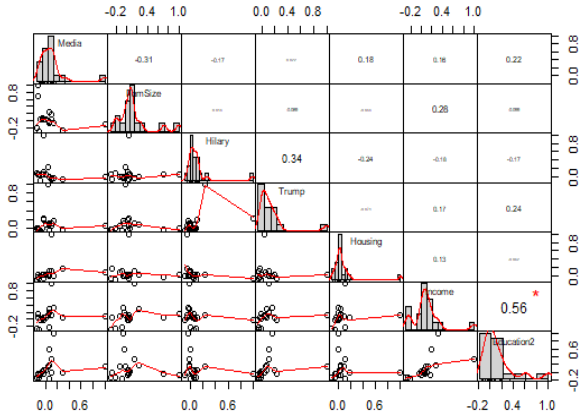
Method 3:

- Now we are fitting the KNN model using the knn() function and which required different inputs like:
  1. A matrix which would be contacting the predictors associated with the training data
  2. A matrix containing the predictors associated with the data for which to make predictions as testing data.
  3. A vector which would be having the training observations.
  4. A value for K, the number of nearest neighbours to be used by the classifier.
- Now we can use the cbind() function to bind the input variables together into matrices and apply knn() function of training set and the test set with proving training observations with the classifier k and K = 3 and which is predicting with 37% accuracy and 41% for K = 10.
- Now we can use the confusion matrix to evaluate the performance of the model. We can see that 51% it has predicted correctly for Democrat, 48% correctly predicted for Republican, 33% correctly predicted for Independent and 2% correctly predicted for others.

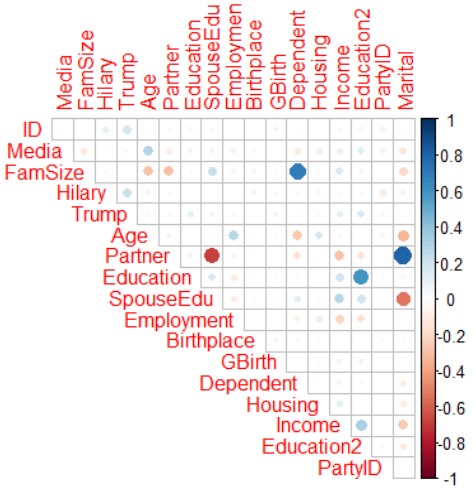
From all the above we see that LDA has the better predictions compared to QDA and KNN.

Results:

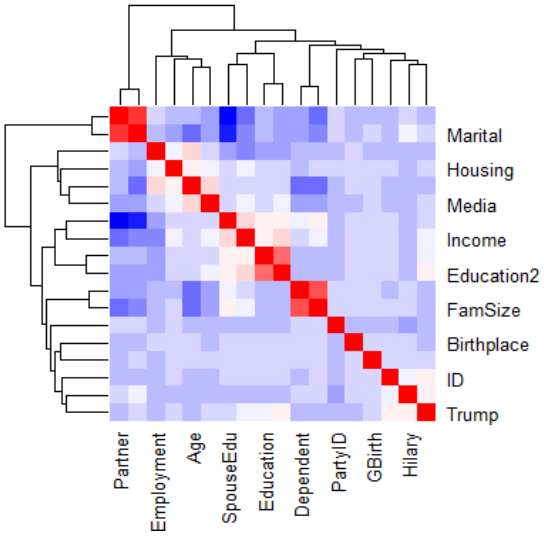
Task 1:
Method 1: Finding the correlation.



From the above plot we can the correlation for the main variables from the data set.



From the above plot we can see the correlation for all the variables in the dataset.



From the above plot we would be see the correlation and the dendrogram.

Method 2: Using Logistic Regression with Trump as the response variable:

- For all the variables in the data set.

```
Call:
glm(formula = Trump ~ ., family = binomial, data = subtrump2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8120  -0.6032  -0.4527  -0.3349   3.7698

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.603e+00  4.897e-01  -3.274  0.00106 **
ID          -1.349e-06  9.546e-07  -1.413  0.15769
Media       -3.230e-02  2.227e-02  -1.451  0.14691
FamSize      3.505e-03  3.994e-02   0.088  0.93006
Hilary       4.856e-01  2.705e-02  17.953  < 2e-16 ***
Age          5.490e-03  3.034e-03   1.810  0.07037 .
Partner     -7.732e-02  5.969e-02  -1.295  0.19515
Education   -9.670e-03  1.589e-02  -0.608  0.54293
SpouseEdu   -2.338e-02  1.008e-02  -2.319  0.02038 *
Employment   7.586e-03  2.019e-02   0.376  0.70716
Birthplace   5.271e-02  3.701e-02   1.424  0.15436
GBirth      -5.605e-02  3.164e-02  -1.772  0.07643 .
Dependent    1.149e-01  5.486e-02   2.095  0.03614 *
Housing     -5.723e-02  3.650e-02  -1.568  0.11693
Income      -1.221e-02  5.430e-03  -2.248  0.02456 *
Education2  -1.714e-01  3.821e-02  -4.486 7.27e-06 ***
PartyID      4.838e-02  4.621e-02   1.047  0.29510
Marital      3.262e-02  4.847e-02   0.673  0.50097
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3993.3  on 4040  degrees of freedom
Residual deviance: 3404.8  on 4023  degrees of freedom
AIC: 3440.8

Number of Fisher Scoring iterations: 5
```

- Using the Logistic Regression and to predict the Trump Variable and using all the other predictors from the dataset. And the confusion matrix will be as follows:

```
> xtabs(~Trump+Pred,data = subtrump2)
            Pred
Trump        Conservative Liberal
  Conservative         3149     102
  Liberal               621     169
>
```

- The mean for the above prediction and its accuracy can be show as below:

```
> mean(subtrump2$Trump==subtrump2$Pred)
[1] 0.8210839
>
```

- Now, we can change the variables for the predicting the Trump variable. And the confusion matrix will be as follows:

```
> xtabs(~Trump+Pred,data = subtrump2)
            Pred
Trump        Conservative Liberal
  Conservative         3151     100
  Liberal               618     172
```

- The mean for the above prediction and its accuracy can be shown as below:

```
> mean(subtrump2$Trump==subtrump2$Pred)
[1] 0.8223212
>
```

Task 2:

Method 1: Linear Discriminant Analysis:

- The confusion matrix for the LDA for all the variables and for the specific variables from the dataset.

```
> table(lda.class, test.data$PartyID)

lda.class   1   2   3   4
        1 217  69 140  14
        2  68 184 105  12
        3  71  59  88   9
        4   1   2   1   1
```

```
> table(lda.class, test.data$PartyID)

lda.class   1   2   3   4
        1 215  68 147  14
        2  68 190 101  13
        3  74  56  86   9
        4   0   0   0   0
```

- The mean for the LDA for all the variables and for the specific variables from the dataset.

```
> mean(lda.class == test.data$PartyID)
[1] 0.4707012
```

```
> mean(lda.class == test.data$PartyID)
[1] 0.4716619
```

Method 2: Quadratic Discriminant Analysis:

- The confusion matrix for the QDA for all the variables and for the specific variables from the dataset.

```
> table(qda.class, test.data$PartyID)

qda.class   1   2   3   4
        1  94  33  68   7
        2 143 205 141  14
        3 119  70 123  14
        4   1   6   2   1
```

```
> table(qda.class, test.data$PartyID)

qda.class   1   2   3   4
        1  59  23  54   2
        2  84 165  88  10
        3 212 123 191  24
        4   2   3   1   0
```

- The mean for the QDA for all the variables and for the specific variables from the dataset.

```
> mean(qda.class==test.data$PartyID)
[1] 0.4063401
```

```
> mean(qda.class==test.data$PartyID)
[1] 0.3986551
```

Method 3: KNN

- The probability of the coefficients with the class k=3

```
> prop.table(xtabs(~test.data$PartyID+knn.pred1),1)
                  knn.pred1
test.data$PartyID          1          2          3          4
                1 0.42016807 0.24929972 0.30812325 0.02240896
                2 0.29617834 0.39808917 0.26751592 0.03821656
                3 0.35928144 0.29041916 0.32634731 0.02395210
                4 0.25000000 0.36111111 0.33333333 0.05555556
```

- The mean for the KNN for K = 3 is:

```
> mean(knn.pred1==test.data$PartyID)
[1] 0.3707973
```

- The probability of the coefficients with the class k=10

```
> prop.table(xtabs(~test.data$PartyID+knn.pred1),1)
                  knn.pred1
test.data$PartyID          1          2          3          4
                1 0.498599440 0.229691877 0.268907563 0.002801120
                2 0.280254777 0.433121019 0.286624204 0.000000000
                3 0.374251497 0.269461078 0.353293413 0.002994012
                4 0.388888889 0.305555556 0.305555556 0.000000000
```

- The mean for the KNN for K = 10 is:

```
> mean(knn.pred1==test.data$PartyID)
[1] 0.4149856
```

Discussion:

Task 1: For the task 1 firstly we have calculated the correlation of the data with respective all the variables in the data and would be checking the best correlation data. Later we would be fitting logistic regression model with keeping Trump as the response variable and all the other variables as the input variables. Next, we would be seeing the highly significance variables from the data and we would be fitting new logistic model with the highly significance variables for the better accuracy.
Task 2: For the task 2, I have fixed LDA, QDA & KNN with all the variables and with preferred variables. And we see that LDA is performing better comparing to all the variables.

References

[1] S. Little, "Select statistics," select statistics, [Online]. Available: https://select-statistics.co.uk/blog/analysing-categorical-data-using-logistic-regression-models/ [Accessed 03 05 2020].

2. Lecture Videos and Lab Videos