**BINARY LOGISTIC REGRESSION**

**INTRODUCTON**:

Binary Logistic regression modelling is done when the dependent variable is categorical and a dichotomous variable. The prediction in logistic regression is based on probability. If the likelihood of the event happening is more than 50% then it is put in one category and the variables for which the odds in favor is less than 50% then it is put in another category. The plot against these two categories forms a ‘S’ shaped sigmoid curve. The equation is represented as:

Log(p(X) / 1 – p(X)) = β0 + β1X1 + β2X2 + ……….+ βpXp

The left hand side is the logit transformation which represents the log of probability of the event happening divided by the probability of the event not happening. The right hand side denotes that one unit increase in X1. changes the log of probability by β1.

**OBJECTIVE:**

The objective of this projec is to check if we can predict wether people has personally experienced discrimination because of their gender in United States. The opinion of about 4867 people will be used for this purpose. We will be chekcing several factors to check if its contributing to the prediction. The final model will contain the features that are significant and contribute to the prediction.

**DATASET** **DESCRIPTION**:

The dataset for the analysis is obtained from the Pew research site – American Trends panel Wave 29 (<https://www.pewresearch.org/american-trends-panel-datasets/>). It consists of survey answers on the topic “Views on Gender” from  Sept. 14 2017 to Sept. 28, 2017. The dataset consists of 4867 rows and 163 columns. Based on the nature of the survey, for this analysis, totally 9 columns has been shortlisted.

**VAIRABLE DESCRIPTION**:

IBM SPSS Statistics 26 software is used for the linear regression analysis.

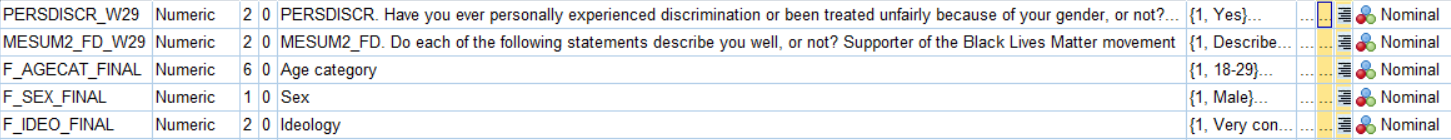


Figure-1 Variable view from SPSS

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Variable Type** | **Dependency** |
| PERDISCR\_W29 (Gender discrimination experience response) | Yes/No | Dichotomous | Dependent |
| Age category | 1 year – 65 years | Categorical | Independent |
| Sex | Male /Female | Categorical | Independent |
| Ideology | Conservative /Non conservative etc. | Categorical | Independent |
| MESUM2\_FD (Black lives matter movement participation) | Participation Yes/No etc. | Categorical | Independent |

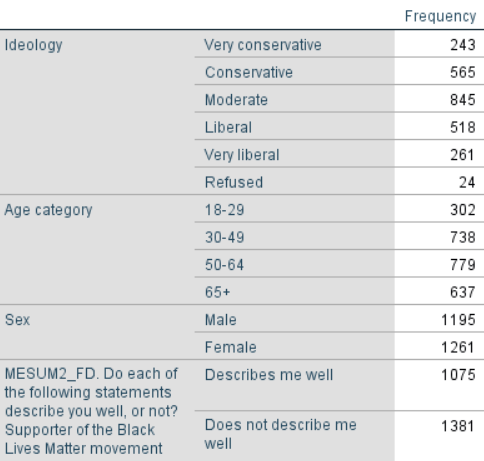


Figure 2- Independent variable categories and their Frequency

**MODEL BUILDING**:

The Dependent variable which is a categorical variable is converted to dichotomous variable by selecting only 2 responses which are Yes and No. The filter is applied in the SPSS.



Several factors which are contribute to the gender discrimination were selected just by knowledge on subject matter. To check if these factors are really contributing to the prediction the significant P value is checked in Block 1 part.

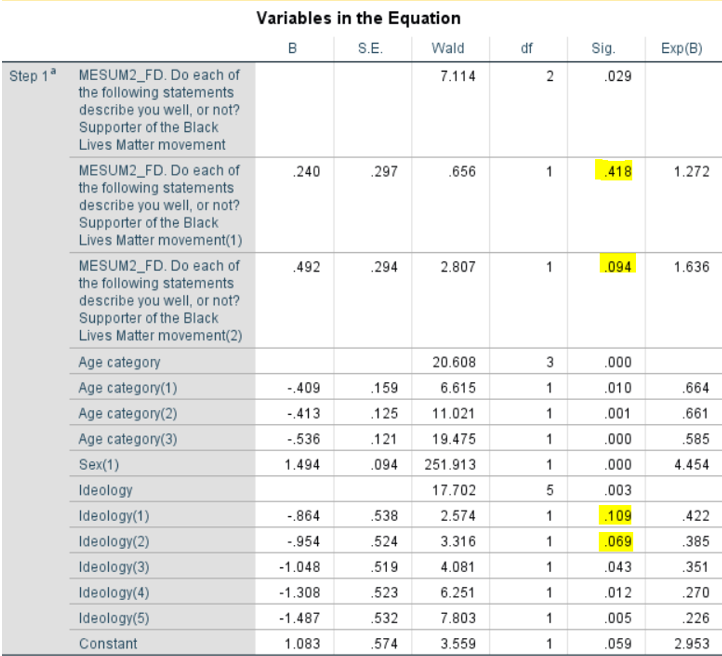


Figure 2- Variables in the equation from Block 1 Method

As we could see from the above figure that categories like MESUM2\_FD(1), MESUM2\_FD(2), Ideology(1), Ideology(2) are not significantly contributing to predicting the gender discrimination as the significance value is greater than 0.05

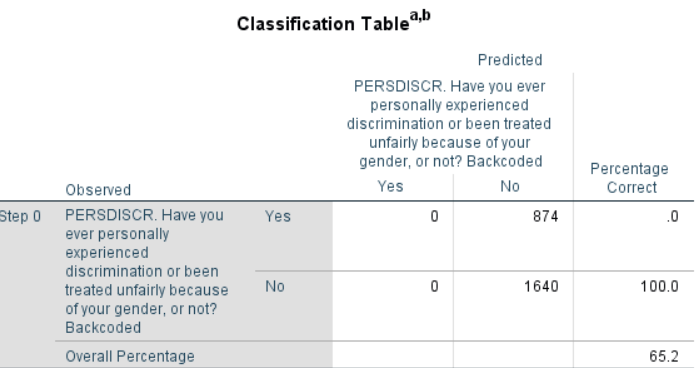
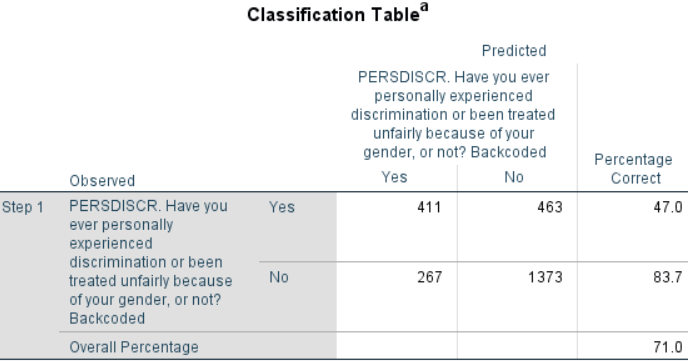
 

Figure 3 – Classification from Block 0 Figure 4 – Classification from Block 1

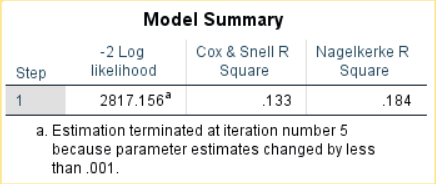
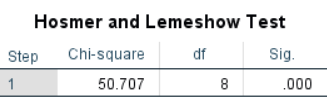
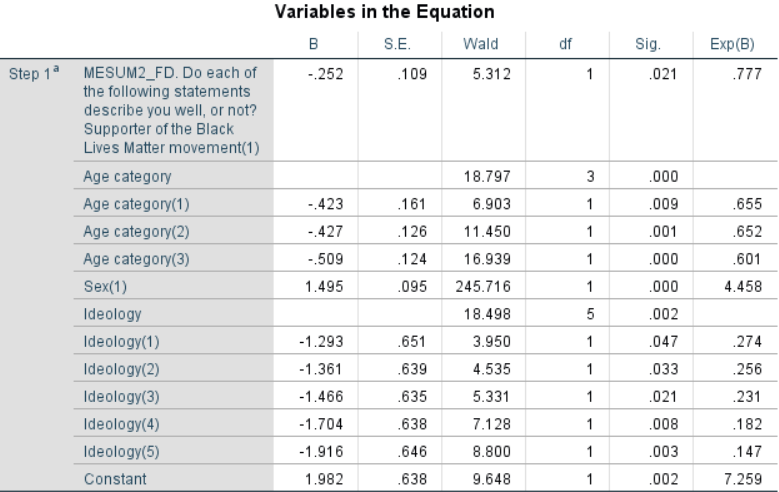
 

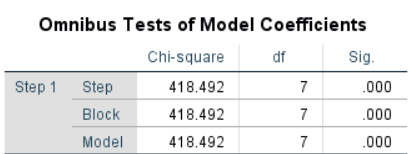
Figure 5 - Model summary from Block1 Method Figure6 – Hosmer and Lemeshow Test

As we could see from the above figure that the PAC (Percentage accuracy of classification) Block 0 is 65.2% and that of Block 1 is 71%. I would be trying to remove the categories which did not significantly contribute to the prediction and the **Pseudo R square values are 18.4%**. The Hosmer and Lemeshow Test shows that the significance value is **0%** which means that the **model is poorly fit**. In order to overcome this the size of the population data will be reduced and a sample data will be used for this analysis. The independent variables with values less than .05 will only be considered.



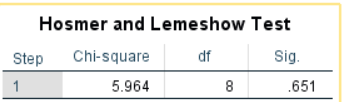
After the categories which is of higher significant values are removed and the sample size has been reduced to 709, below is the significant P values of the rest of the independent categorical variables. As we could observe the significance value is below 0.05 for all the independent categorical variables. Hence, we could understand that all of the independent variable categories are significantly contributing to the prediction of dependent variable.

Omnibus Tests for model coefficients:

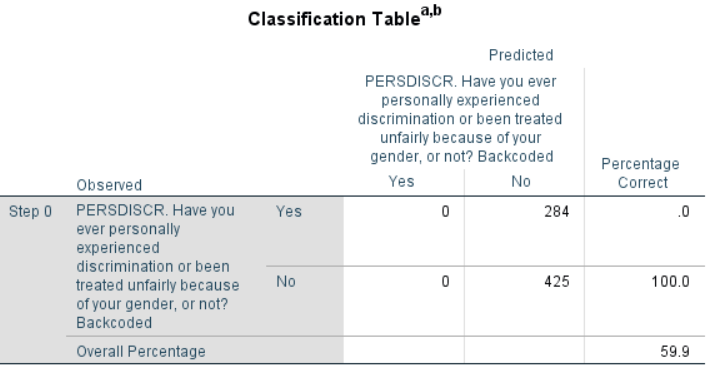
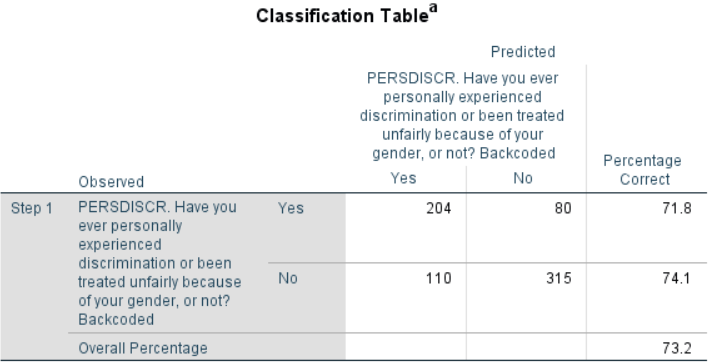


The Omnibus Tests of Model coefficients are used to check how the new model performs against the baseline or Block 0 model. The significance value of 0.000 states that our new model is performing better than the baseline model.

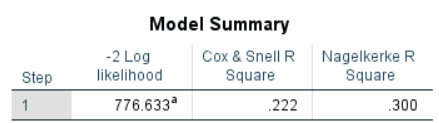
Hosmer and Lemeshow Test



Hosmer and Lemeshow Test is performed how well the model is fit. The model would be a poor fit if the significance value is below 0.05. The value 0.9 indicates support for the model. The significance value 0.651 states that the variables provide a moderate support for the model. Since it is not a best practice to standardize categorical variables, the sample data should be wisely chosen based on the domain knowledge of the dataset to improve the Significance value in Hosmer and Lemeshow Test.

The Block 0 states that the maximum response from the people is that they have not experienced gender discrimination. Hence the model evaluated all the response to ‘No’, in that case the PAC is 59.9%. If we check the Block 1 after the contribution of the independent variables the PAC has improved to 73.2% which states the model was able to classify 73.2% of the cases correctly.



The -2Log likelihood is similar to standard error and the accuracy of the model increases with lesser -2LL value. The Nagelkerke R Square value 30% represents the amount of variation in the dependent variable that has been predicted by the independent variable.

Check for Assumptions:

Dependent variable outcome:

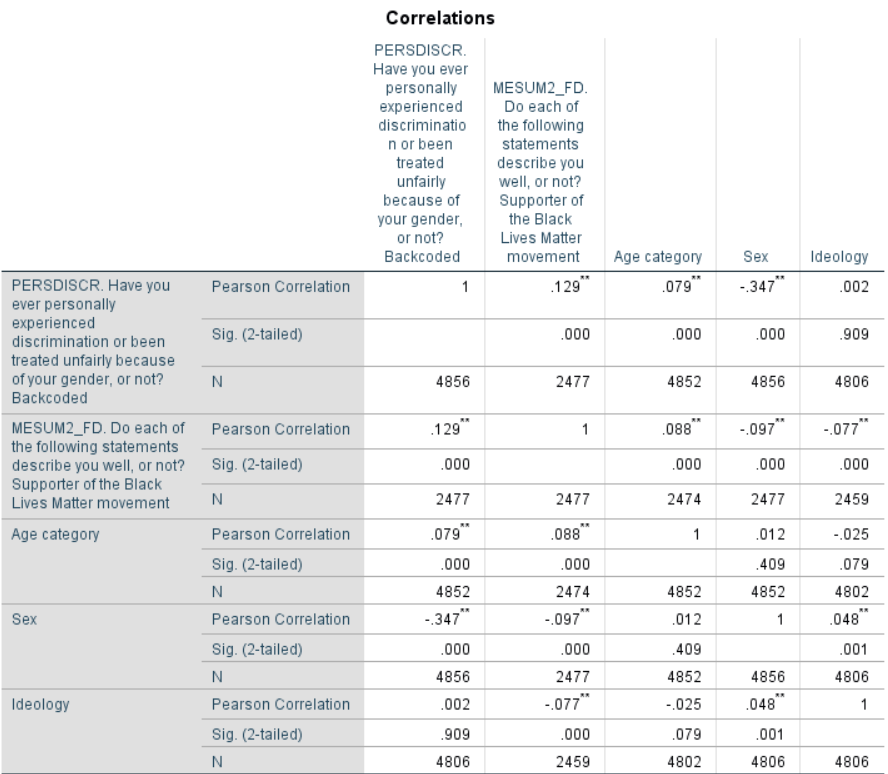
The dependent variable – Whether people has experienced gender discrimination takes two responses YES(1) and NO(2), these two events cannot happen simultaneously. Hence the dependent variable outcome is mutually exclusive.

Sample size:

Binary logistic regression tends to work better with large data. The minimum criteria is to have at least 20 instances for an independent variable. In this analysis we saw 709 samples, hence the assumption is satisfied.

Multicollinearity:

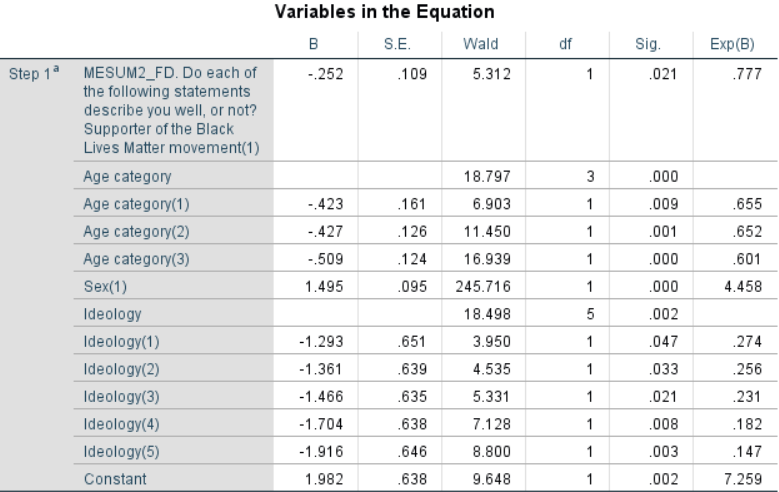
The correlation among independent variables results in a model with poor accuracy. Hence the correlation should be minimal. In order to check this assumption, Pearson’s correlation will be used. It’s a bivariate correlation check to find out correlation among two independent variables. Ideally for Pearson’s correlation, the value should not exceed 0.7, as we could see the correlation among the independent variables has not exceeded 0.7 (The value 1 states a complete correlation). Hence the assumption is satisfied.



Model summary and conclusion:

The logistic regression equation for this analysis can be represented as below

E(y)=e1.98-0.25(x1)-0.42(x2)-0.42(x3)-0.50(x4)+1.5(x5)-1.29(x6)-1.36(x7)-1.4(x8)-1.7(x9)-1.9(x10) /(1+e1.98-0.25(x1)-0.42(x2)-0.42(x3)-0.50(x4)+1.5(x5)-1.29(x6)-1.36(x7)-1.4(x8)-1.7(x9)-1.9(x10)



* The people who have a “conservative” Ideology (Ideology 1) are e-1.36 times less likely to experience gender discrimination than people who have “very conservative”,” Moderate”, “Liberal” ideologies by a factor of 0.25.
* The Exp(B) column represents the values which increase the odds in favor or likelihood of the even happening. From the table, the category Sex has the highest value, which can be stated as, Sex of the person is highly influential in predicting the dependent variable outcome and it increases the odds in favor of a person experiencing gender discrimination by a factor of 4.45.