Statistics for Data Analytics
Submitted to: Prof. Tony Delaney

Prediction of the Election Outcome by using a Binary Logistic Regression Model

Jaswinder Singh[1]*

[1] *x19219997@student.ncirl.ie*
***Affiliation**: *Cohort 'B'*, MSc in Data Analytics, National College of Ireland, Dublin, Ireland

Abstract

This part of the aims to predict the outcome of an election by deploying a binary logistic regression model. The data used for the analysis was taken from one of the surveys published on the Pew Research centre webpage. The model is evaluated on the basis of *Pseudo R squared values*(Cox & Snell and Nagelkerke), *Wald's test* and *Hosmer and Lemeshow* goodness of fit test. The final model is evaluated by using various parameters like accuracy(calculated from the confusion matrix), ROC(Receiver Operating Characteristic) curve and AUC(Area Under Curve). The analysis was done using IBM SPSS software and R.

1. INTRODUCTION

Today, the mainstream media captures every little movement of the political leaders, analyze them and often attempt to predict the most likely step going to be taken by them. The media coverage during the presidential elections is so enormous that it triumphs every other issue going on at that time. Now, every popular news blog and channel tries to predict the election outcomes way before the actual results come out. These predictions are often based on the surveys conducted by their trusted sources and often include large number of people giving their opinions on a particular subject/ question asked by the reporter. After collecting the data from hundreds of thousands of people, the analysis is carried out using various statistical techniques and models and then the results are published. In this report, we also aim to predict the election outcomes by employing a binary logistic regression model using various predictor variables. We then use the confusion matrix obtained to calculate the efficiency and accuracy of the model. By using different combinations of the predictor variables, we try to increase the accuracy of the model(simultaneously analyzing other parameters as well) while also minimizing the false positive and false negative outcomes.

2. BINARY LOGISTIC REGRESSION - A BRIEF INTRODUCTION

The binary logistic regression is a regression technique in which the response variable is dichotomous in nature i.e it has only two possible values. The predictor variables could be of any type - numerical or categorical. In the multinomial logistic regression, on the other hand, the response variable can have more than two categories. The logistic regression can be used to predict the probability of an event occurring on the basis of one or more predictor variables. It can also be used to calculate the *odds ratio* which tells us the influence of a predictor variable on a response variable. Mathematically, the relationship between the predictor variables(say x_1, x_2, x_3,..x_p) and the response variable y can be described by the equation:

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \tag{1}$$

Here $E(y)$ cna be interpreted as the probability of y = 1(If 2 possible values of y are 0 and 1) given x_1, x_2, x_3,..x_p. A simple random sample is then used to compute the coefficients b_0, b_1,...b_p which are the point estimators of β_0, β_1,...β_p.

In the linear regression, the parameters b_0, b_1,...b_p are estimated using the Least- Squares estimation, while in binary logistic regression, these parameters are estimated on the basis of $loglikelihood function$. The parameter estimates are then tweaked until the best value for the log likelihood function is obtained.

Logit Transformation: As we know that the odds of of an event(in favor) which has a probability of happening as $\hat{p}$, can be given as:

$$Odds = \frac{\hat{p}}{1-\hat{p}} \tag{2}$$

Hence we can write the equation for logistic regression as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+..+\beta_p X_p \tag{3}$$

The term $\log\left(\frac{p(X)}{1-p(X)}\right)$ is called the *log-odds* or *logit*. Hence changing X by one unit changes the log odds by β_1.

The odds ratio can be defined as the ratio of change in odds after one unit change in the predictor variable to the original odds.

$$Odds = \frac{\text{Odds after a one unit change in predictor variable}}{\text{Original Odds}}. \tag{4}$$

The odds ratio actually tells us the change in the odds of one of the categories of the response variable resulting from a unit change in the predictor variable.

2.1 Assumptions of Logistic Regression

The following assumptions should be met for a binary logistic regression model:

1. The predictor variables should not be highly correlated with each other but should be strongly correlated to the response variable.
2. The outcomes for the response variable should be mutually excusive.
3. The sample size should be large enough(20 cases per predictor).

2.2 Tests for evaluating a binary logistic regression model

1. *Wald's t-statistic Test*: The wald's test tests the null hypothesis(H_0) that $B = 0$. The p- values of the Wald statistic of the predictor variables tell us which of our variables are statistically significant and should be included in the final model.
2. *Omnibus test for the coefficients:* The null hypothesis is that the coefficients of all predictor variables are zero. This is tested by a χ^2 statistic with degrees of freedom equal to the number of predictor variables. The p-values should be significant(< 0.05)
3. *Pseudo* R^2: There are two pseudo R^2 values that we will consider: Cox & Snell, Nagelkerke. The maximum value is 1(similar to R^2 in multiple regression.
4. *Hosmer and Lemeshow test*: The Hosmer and Lemeshow tests for the quality of fit for the model. The poor fit is indicated by the p-value of less than 0.05.
5. *Log likelihood* : Lower the value of log likelihood, better is the fit.

3. DATA CLEANING AND PRE PROCESSING

The data used for the purpose of this project was compiled from the Pew Research Centre survey database. The data was actually from a 2016 survey done at the time of the presidential elections to determine whether people would support Donald Trump in the upcoming elections or not. The questions in the survey were converted into various predictor variables and were named accordingly. The data was cleaned and prepared for further analysis using R. Some tests were done using SPSS while the overall analysis was done using R.

4. MODEL ANALYSIS

4.1 Description of variables used

The data contains the following predictor(independent) and response(dependent) variables:

- **Dependent/Response Variable**:
  1. Support Trump(0 or 1)
- **Independent Variables**: Figure 1 explains the types and values of the predictor variables used in the model.

Independent Variable	Type	Values
Age_Category	Categorical	18-29, 30-49, 50-64, 65+
Sex	Categorical	Female, Male
Education_Category	Categorical	'Some College', 'No degree', 'Associate degree', 'College graduate', 'Postgraduate', 'High School graduate', 'Less than high school'
Hispanic	Categorical	'Yes', 'No'
Combining_Race	Categorical	'White',' Asian or Asian American', Black or African American', 'Some other race', 'Mixed race'
Race_Ethnicity	Categorical	'Hispanic', 'White non Hispanic', 'Black non Hispanic', 'Other'
Marital_Status	Categorical	'Divorced', 'Married', 'Widowed', 'Never been married', 'Living with a partner', 'Separated'
Religion	Categorical	'Roman Catholic', 'Protestant', 'Jewish Agnostic' (not sure if there is a God), 'Atheist (do not believe in God)', 'Mormon', 'Orthodox(VOL)', 'Unitarian (Universalist)(VOL) Christian', 'Muslim', 'Buddhist', 'Hindu'
Party	Categorical	'Republican', 'Democrat', 'Independent'
Income_Recode	Categorical	'20 to under $30,000', '50 to under $75,000', '100 to under $150,000', '75 to under $100,000', '$150,000 or more', '30 to under $40,000', '10 to under $20,000', 'Less than $10,000', '40 to under $50,000'
Ideology	Categorical	'Moderate', 'Conservative', 'Very liberal', 'Liberal', 'Very conservative'
Internet_frequency	Categorical	'Every day', 'At least once a week but not every day', 'Never', 'Once a month', 'Once a week', 'Less than once a month'
Social_media_user	Categorical	'Social Media Users', 'Not Social Media Users'

Figure 1. Description of variables used

4.2 Visualisation of predictors with the response variable

The following figures(Figure 2 to 9) show the proportion of the each of the categorical values of the dependent variable(0 and 1) shared by each category of the predictor variables. Since all the predictors are categorical, we cannot create a correlation diagram but we can still visualize the correlations by using the mosiac plot available in R. Figure 10 illustrates a mosiac plot between the dependent variable $Support_Trumpandindependentvariables Sexand Party$.

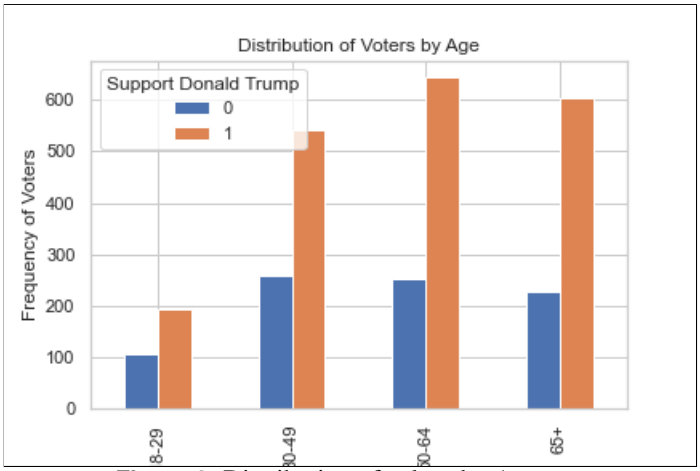

Figure 2. Distribution of values by Age

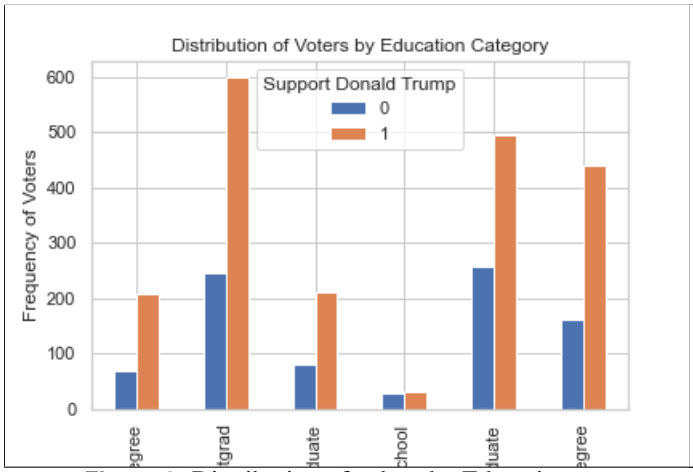

Figure 3. Distribution of values by Education

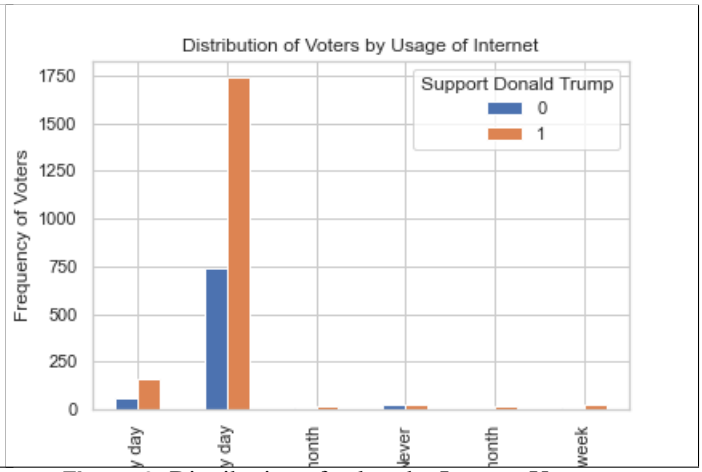

Figure 6. Distribution of values by Internet Usage

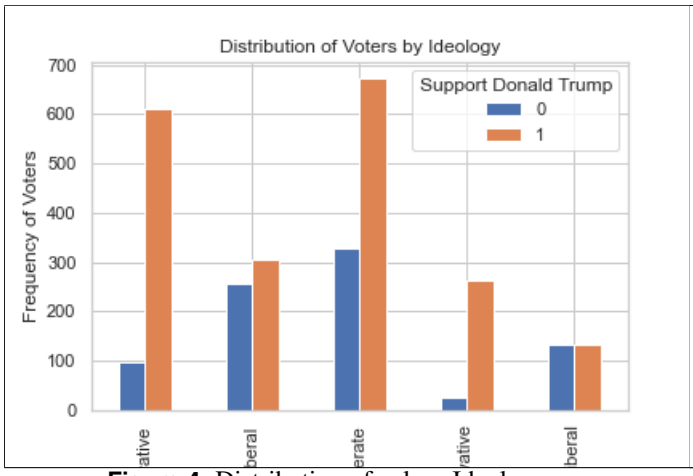

Figure 4. Distribution of values Ideology

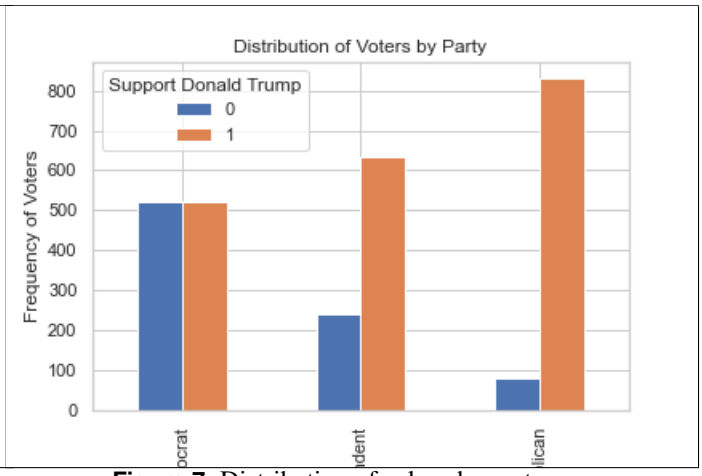

Figure 7. Distribution of values by party

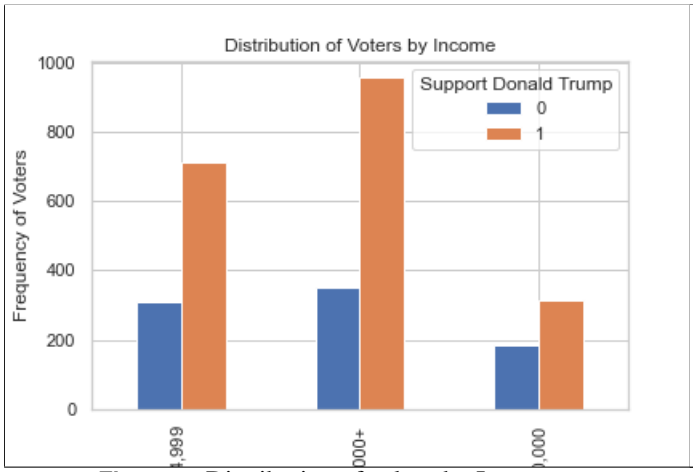

Figure 5. Distribution of values by Income

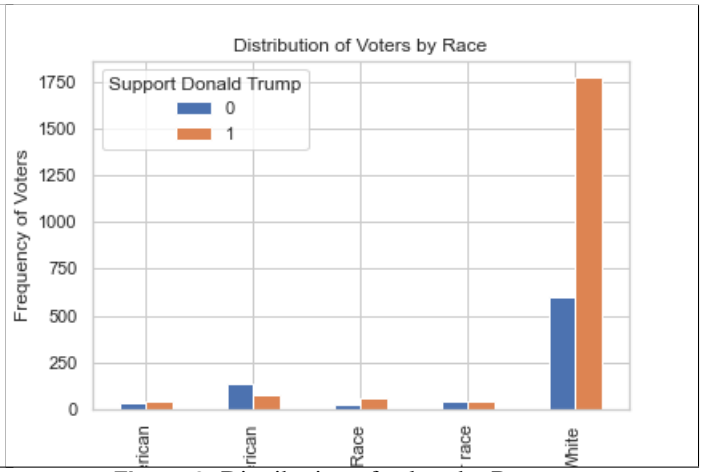

Figure 8. Distribution of values by Race

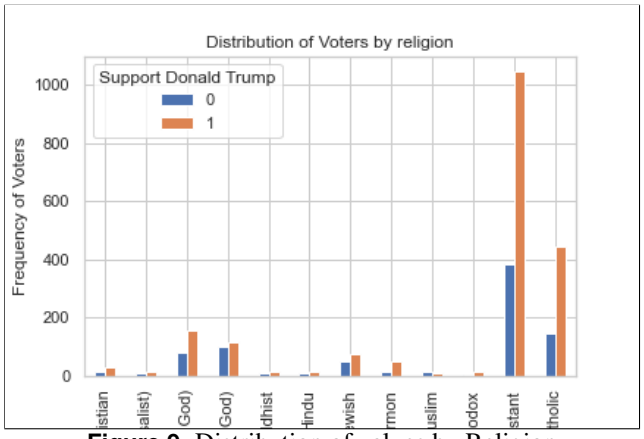

Figure 9. Distribution of values by Religion

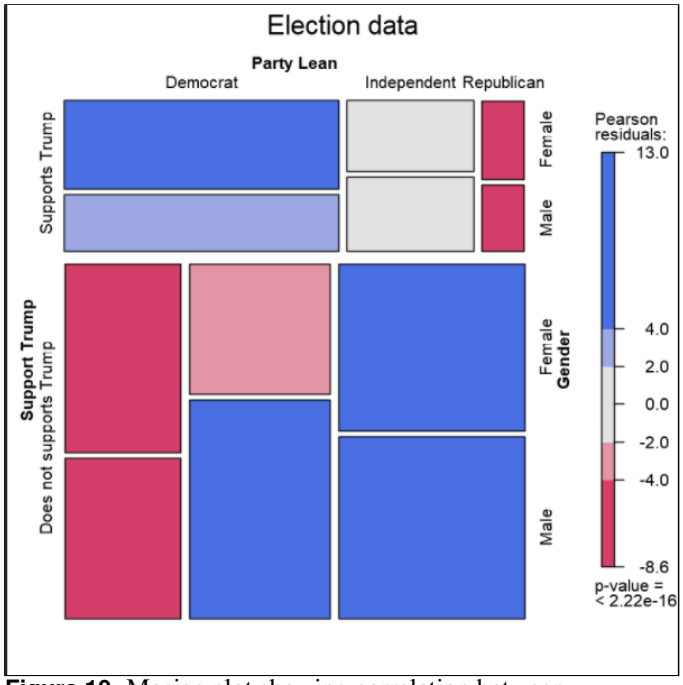

Figure 10. Mosiac plot showing correlation between different categorical variables

4.3 Modelling

The following model was selected as the final model after evaluating a few models on the basis of metrics discussed in section 2.

- ***Dependent Variable***
  1. *Support Trump*
- ***Independent Variables***
  1. *Sex*
  2. *Education Category*
  3. *Party*
  4. *Income Recode*
  5. *Combining Race*
  6. *Ideology*
- ***p-values****: Refer to figure 11 and 21*
- ***Tests****: The tests are shown in following figures:*
  - *Hosmer and Lemeshow test : Figure 12*
  - *Omnibus χ^2 test : Figure 13*
  - *Pseudo R^2 : Figure 14*
  - *Wald's test : Figure 15*
  - *Anova χ^2 test: Figure 16*
- ***AIC Value*** *: Figure 17*
- ***Confusion Matrix****: Figure 18*
- ***Efficiency*** *: Figure 19*
- ***ROC Curve*** *: Figure 20*
- ***AUC Value*** *: Figure 20*

	Pr(>|z|)
(Intercept)	0.49596
SexMale	0.00019 ***
Education_CategoryCollege graduate/some postgrad	0.21449
Education_CategoryHigh school graduate	0.23042
Education_CategoryLess than high school	0.00955 **
Education_CategoryPostgraduate	0.04685 *
Education_CategorySome college, no degree	0.99061
PartyIndependent	2.43e-06 ***
PartyRepublican	< 2e-16 ***
Income_Recode$75,000+	0.36371
Income_Recode<$30,000	0.29594
Combining_RaceBlack or African-American	0.04194 *
Combining_RaceMixed Race	0.12056
Combining_RaceOr some other race	0.50480
Combining_RaceWhite	0.00150 **
IdeologyLiberal	1.47e-06 ***
IdeologyModerate	5.04e-05 ***
IdeologyVery conservative	0.13968
IdeologyVery liberal	4.59e-07 ***

Figure 11. p- values for coefficients

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.942	8	.543

Figure 12. Hosmer and Lemeshow test

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	583.069	15	.000
	Block	583.069	15	.000
	Model	583.069	15	.000

Figure 13. omnibus test for coefficients

CoxSnell: 0.187645238207481
Nagelkerke: 0.266305715607579

Figure 14. Pseudo R^2 Values

```
Wald test for Sex Education_Category Party Income_Recode Combining_Race Ideology:Support_Trump
 in glm(formula = Support_Trump ~ Sex + Education_Category + Party +
    Income_Recode + Combining_Race + Ideology, family = binomial(link = "logit"),
    data = elec_df)
F =  18.60938  on  14  and  2810  df: p= < 2.22e-16
```

Figure 15. Wald's test for coefficients

A anova: 7 × 5

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
	<int>	<dbl>	<int>	<dbl>	<dbl>
NULL	NA	NA	2828	3449.977	NA
Sex	1	27.127561	2827	3422.850	1.904631e-07
Education_Category	5	23.375381	2822	3399.474	2.861048e-04
Party	2	403.946779	2820	2995.527	1.923409e-88
Income_Recode	2	9.686485	2818	2985.841	7.881457e-03
Combining_Race	4	80.375746	2814	2905.465	1.450100e-16
Ideology	4	43.405553	2810	2862.060	8.524774e-09

Figure 16. ANOVA test

```
LR_Model2$aic
```

```
2900.05964876738
```

Figure 17. AIC Value for the model

	Predicted Negative	Predicted Positive
Obseved Negative	319	526
Observed Positive	200	1784

Figure 18. Confusion matrix for the model

```
efficiency2 <- sum(diag(mytable2))/sum(mytable2)
efficiency2
```

```
0.743372216330859
```

Figure 19. Efficiency for the model

```
Data: LR_Model2$fitted.values in 845 controls (Support_Trump 0) < 1984 cases (Support_Trump 1).
Area under the curve: 0.775
```

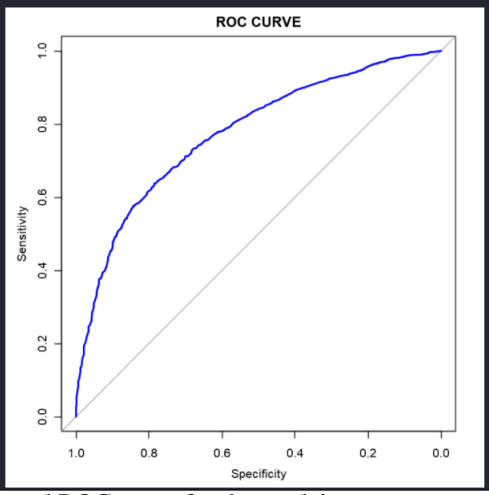

Figure 20. AUC value and ROC curve for the model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Sex(1)	-.353	.094	14.184	1	.000	.703	.585	.844
	Education_Category			9.169	2	.010			
	Education_Category(1)	-.291	.112	6.821	1	.009	.747	.600	.930
	Education_Category(2)	-.383	.159	5.763	1	.016	.682	.499	.932
	Party			92.508	2	.000			
	Party(1)	-1.535	.160	92.062	1	.000	.215	.157	.295
	Party(2)	-.998	.150	44.035	1	.000	.369	.274	.495
	Ideology			42.423	4	.000			
	Ideology(1)	.987	.193	26.096	1	.000	2.684	1.838	3.920
	Ideology(2)	.170	.154	1.211	1	.271	1.185	.876	1.604
	Ideology(3)	.393	.155	6.423	1	.011	1.482	1.093	2.008
	Ideology(4)	1.344	.270	24.758	1	.000	3.835	2.258	6.511
	Income_Recode			2.782	2	.249			
	Income_Recode(1)	-.232	.139	2.782	1	.095	.793	.604	1.041
	Income_Recode(2)	-.078	.106	.535	1	.465	.925	.751	1.140
	Combining_Race			90.350	4	.000			
	Combining_Race(1)	-.848	.258	10.775	1	.001	.428	.258	.711
	Combining_Race(2)	-1.419	.167	72.011	1	.000	.242	.174	.336
	Combining_Race(3)	-.269	.251	1.149	1	.284	.764	.467	1.250
	Combining_Race(4)	-1.090	.252	18.782	1	.000	.336	.205	.550
	Constant	2.040	.231	77.684	1	.000	7.687		

a. Variable(s) entered on step 1: Sex, Education_Category, Party, Ideology, Income_Recode, Combining_Race.

Figure 21. Summary of Coefficients for the model

5. Conclusion

As we have seen in the previous series of figures, our model fits the data good enough. The coefficients shown in the figure 21 can be used to construct an equation for the model if desired. From the ANOVA table The p-values of all the predictors used in the model are significant(Figure 16). The values of efficiency(Figure 19) and the AUC(Figure 20) are also good, if not optimum. Although the pseudo R^2 values are not optimum(figure 14), but all the other tests show that the model is a good fit to the data. The missing values for Exp(B) in figure 21 is because in each predictor variable, one category is chosen as the reference category, and the B and Exp(B) values for all the other categories are relative to that reference category.