STAT30270 – STATISTICAL MACHINE LEARNING (OL)

STUDENT NO – 18201501

ASSIGNMENT – 3

Introduction:

Titanic data set consist of 1313 observations and 11 variables. It literally means, there were 1313 passenger's travelled in titanic boat categorized based on the tabulation's mentioned below in fig1. Moreover, the aim of our model is to predict survival of the passengers of the boat using logistic regression technique and to interpret the performance of the model.

Passenger class	Home destination
Survived	Room
Name	Ticket
Age	Boat
Embarked	Sex

Figure 1. Categorization of passengers travelled in titanic boat

Data Modelling:

From the dataset it is evident, most of the passenger's age are updated with NA values. In order to predict the survival of passengers with the input observations the missing values need to replace with median values. Hence, made use of the random forest classification for updating the missing values in the dataset. Therefore, once the missing values are replaced we go ahead with fitting the model using logistic regression. Additionally, modified the categorical data using dummy variables using factors.

Fitting using Logistic Regression:

GLM function in R is used to perform two class logistic regression. Firstly, performed with all the predictor values for which the algorithm did not converge properly. Secondly, build the model only the most significant predictor variable as mentioned below.

```
Call:
glm(formula = survived ~ pclass + titanic.age + sex, family = "binomial",
    data = imputed_titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0132  -0.6465  -0.3418   0.5764   2.7676

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.662622   0.394245  11.827  < 2e-16 ***
pclass2nd   -1.627417   0.232768  -6.992 2.72e-12 ***
pclass3rd   -3.017156   0.227184 -13.281  < 2e-16 ***
titanic.age -0.066845   0.008187  -8.164 3.23e-16 ***
sexmale     -2.445220   0.157588 -15.517  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1686.8  on 1312  degrees of freedom
Residual deviance: 1124.4  on 1308  degrees of freedom
AIC: 1134.4

Number of Fisher Scoring iterations: 5
```

From the fitted model using the coefficient of the significant variables, we infer the odds value of the predictor variables passenger class, age and sex are most significant in predicting the survival rate of the passengers. From the summary of the logistic model is seen that on one hand, passengers travelling in class1 have higher survival rate compared to passengers travelling in class2 and class3. On other hand, male passengers have low survival rate of 2.445 which means that female have higher survival rate comparatively. Also, on comparing the survival with respect to age it is predicted

that as the age of the passengers increase there is decrease in their possibility of survival as estimated coefficient is -0.066.

Also, from the model seen in figure2, it seen that the null deviance value of logistic model is 1686.8 without including any of the predictor variables and residual deviance is 1124.4 which includes all the significant predictor variables. The model has a difference of 12 parameters and difference in deviance of 562. Since the Chi-square value with 4 degree of freedom is greater than 562. The deviance with saturated model is good comparative to deviance with no predictor variables.

Residuals:

Deviance residuals yields a type of residuals to check the correctness of the analysis. From the residual deviance plot in figure3 it seen the regression line is almost flat as the abline starting from the curve 0 and ending at curve 1 covering most of the observation which results in minimal probability of outliers in the model. Hence, the model is proven to be good over residual deviance.

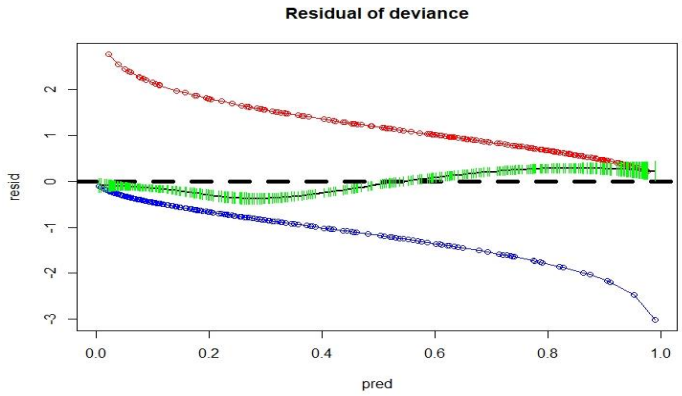

Figure3. Residual of deviance

Goodness-of-fit:

Goodness-of-fit reveals comparison between the observed value of response variable to the expected value of fitted model. Methods of assessing the goodness-of-fits are Pearson test, deviance test and HL test. Evaluating the goodness-of-fit using Pearson test reveals the below inferences.

```
Pearson goodness-of-fit test for logistic regression models
with normal approximation

Chi-square statistic:  1442.062
Normal deviate 0.570328
P-value:  0.5684552
```

Pearson test results the chi-squared statistics of 1442.06 with p-value of 0.5. Since the p-value is above the significant value of 0.05, the model fit is acceptable.

Performance measure:

Performance measure is used in evaluating the consistency measure between two measures. The two most popular measure is accuracy and Area under ROC curve. ROC curve is plotted with False positive value (FP) against True positive value (TP) as shown in the figure4. From the ROC plot, it is inferred that the ROC curve is similar to ideal one with slight deviation True positive point of 0.7. Hence, the model has good performance measure.

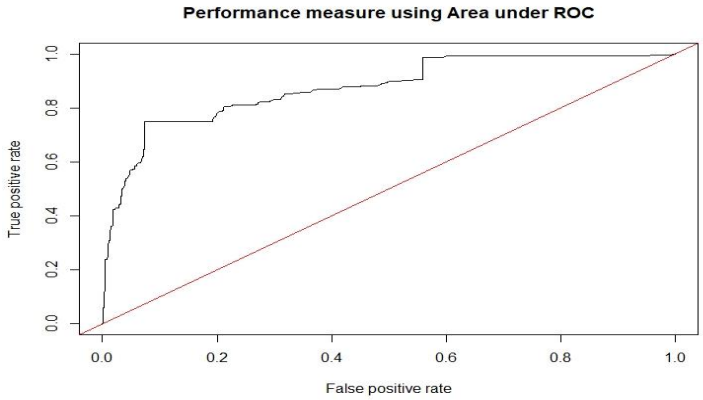

Figure4. Area under ROC

Additionally, made use of the "SDMTools" package in R to determine the optimum threshold value to predict the rate of survival of the passengers. Using, which the average optimum threshold value is found to 0.46 as shown below in figure5, which literally means that the observations above the threshold value of 0.46 is predicted to be survived and the observation value below 0.46 is predicted to be one who failed to survive.

```
$min.ROC.plot.distance
[1] 0.46
```

Figure 5. Optimum threshold

```
> roc_obj <- roc(survived, predicted_survival)
> auc(roc_obj)
Area under the curve: 0.7918
```

Figure 6. AUC value

Accuracy of the model explains how well the performance measure is classified. Accuracy is calculated using the cross tabulation table

		Predicted	
		0	1
Truth	0	True Negative (*TN*)	False Positive (*FP*)
	1	False Negative (*FN*)	True Positive (*TP*)

```
> table(survived,predicted_survival)
        predicted_survival
survived   0   1
       0 689 175
       1  96 353
```

From the cross tabulation, its found that 689 passengers who did not survive were correctly predicted, whereas 175 passengers who did not survive were wrongly predicted as survived. On the other hand, 353 passengers who survived were correctly predicted to be survived whereas 96 passengers who survived where wrongly predicted as they failed to survive. Hence, the accuracy of the model is 0.79.

Conclusion:

- From the summary of the model it is inferred that, better the passenger class higher the survival rate.
- Also, women have higher survival rate than male.
- The model fitted with logistic regression is good as the accuracy of its survival prediction is 79 percentage.
- Accuracy of the model is 0.79%, hence the misclassification rate is minimum with 0.21%.
- Good fit of the model is proved using the Pearson goodness-of-fit.
- Additionally, the Area under the curve calculated using the ROC is 0.791 which represents that the logistic regression model holds good in predicting the survival of passengers, since higher the AUC value better the model.