COMP 237 - Online lab assignment “Logistic Regression”

1. **Exercise # 1:** titanic analysis and logistic regression (100 marks)
2. Initial Exploration:

**4) Identify four columns that are not going to be useful for the model.**

Ans: ‘Ticket’,’PassengerId’,’Cabin’, ’Name’

Reason: -

1. ‘Ticket’ – This column contains alphabetic and numeric each value moreover it scores of distinctive values.
2. ‘PassengerId’- All values are distinctive during this column and thus this column may not facilitate in prediction and index of data frame do the duty of all distinctive values in column.
3. ‘Cabin’ – This column has 687 missing values. In machine learning, we can predict additional accurately we tend to have additional information, currently the column that cannot give information which are less helpful.
4. ‘Name’ – In logistical regression, we must convert information of columns in numerical and conjointly if there’s categorical information in column then we tend to try and convert data in small number of categories. Both requirements are hard to fulfill for this column and that is why dropped this column
5. Data visualization:

**1.(C) Analyze both plots and write a conclusion from each plot in your written response.**

A bar graph showing the # of survived versus the traveler category –

1. It may be seen that highest range of passengers could not survived at school three.
2. Successful survival rate is nearly same for all three categories. However, in class 1, slightly additional range of passengers survived.

A bar graph showing the # of survived versus the gender: -

1. It is evident from the bar graph that prime range of male traveler (more than four hundred) could not survived compared to feminine passenger.
2. Same manner, over two hundred feminine passengers were survived that is kind of over male passengers.
3. **Analyze the output and write some conclusions for scatter matrix.**
4. PClass – survived matrix = PClass has categories between 1,2,3 and survived has category 0 and 1 and hence, are showing on 6 points. Where (0,1), (0,2),(0,3), (1,1), (1,2), (1,3)
5. Fare-Survived matrix = values in Fare column is distributed almost same between 0 and 1 value of survived column that mean Yes or No or survival
6. PClass-Fare = Pclass has categories and Fare column has many values. Matrix of both shows many values Fare situated are 3 points(categories) of PClass
7. Data transformation (round #1):

**12. Form the histogram generated focus on the “Port of Embarkation” and write in your written response some highlights**

Histogram shows that, highest number of passengers were from S = (Southampton). On other hand, lowest number of passengers were from Q = Queenstown. And, C = Cherbourg is lying in between.

1. Build a model

3 Cross validation:

* + - 1. **Note these results in your writer report and recommended the best split scenario.**

**--------------------------------------------------------------------------------------------------------------------------------**

10.0 % = 0.7125 0.7965277777777777 0.85

15.0 % = 0.7368421052631579 0.8005789473684211 0.868421052631579

20.0 % = 0.7222222222222222 0.7964593114241001 0.8450704225352113

25.0 % = 0.746268656716418 0.7950022614201718 0.8507462686567164

30.0 % = 0.746031746031746 0.7978238607270864 0.8387096774193549

35.0 % = 0.7241379310344828 0.8013914095583787 0.8620689655172413

40.0 % = 0.7407407407407407 0.7979035639412997 0.8333333333333334

45.0 % = 0.7551020408163265 0.8040816326530612 0.8571428571428571

50.0 % = 0.7111111111111111 0.791212121212121 0.8636363636363636

**--------------------------------------------------------------------------------------------------------------------------------**

Middle column in above table, shows mean score for different percentage of testing set. I kept more and less suitable percentage test set in bold.

We can see that, when we keep test data 45% ,cross validation gived highest score(Best split scenario )and on other hand, for 25%, score is lowest.

* 1. Test the model

1. **Write down and note the values of : accuracy, precision and re-call**

**When Keeping threshold to 50% = 0.5**

**Accuracy = 0.7910447761194029**

**precision recall**

0.0 0.85 0.82

1.0 0.71 0.75

**-------------------------------------------------------------------------------------**

**When Keeping threshold to 75% = 0.75**

**Accuracy = 0.7835820895522388**

**precision** **recall**

0.0 0.75 0.98

1.0 0.94 0.45

-----------------------------------------------------------------------

1. **Compare the values of accuracy, precision and re-call generated at the threshold 0.5 and 0.75 .**

* Precision for 0 is more when threshold is 75% compared to 50% threshold.
* Same way, recall is also high when threshould is 75% compared to 50% threshold.

**Graphical Data : -**





