Logistic Regression

Statistical Prediction Modeling *DATA 2204*

Assignment No1

Due Date : ,17th Feb 2022

Jumana Obeid

100832624

DURHAM COLLEGE
SUCCESS MATTERS

b. Rational Statement (summary of the problem or problems to be addressed by the PPT) – 2%

Mr.Hughes, In this problem, needs to predict if the patient have cancer based on some input, the algorithm that we will use is "Logistic regression" this algorithm will return the probability of a having either a Benign or Malignant cancer.

This type of analysis can help you predict the likelihood of an event happening or a choice being made.This algorithm assumes that the output , follow a binomial distribution, and that there is a linear relationship between the independent variables and the link function(logit).Also , the independent variables(input/features) have mutually exclusive and exhaustive categories.The dataset contains the following variables:

ID - ID number

1. Clump Thickness – 1-10
2. UofCSize - Uniformity of Cell Size 1-10
3. UofShape - Uniformity of Cell Shape 1-10
4. Marginal Adhesion - 1-10
5. SECSize - Single Epithelial Cell Size 1-10
6. Bare Nuclei - 1-10
7. Bland Chromatin - 1-10
8. Normal Nucleoli - 1-10
9. Mitoses - 1-10

Dependent Variable Class –

1. Benign (i.e. No Cancer) - 2,
2. Malignant (i.e. Cancer) - 4

DURHAM COLLEGE
SUCCESS MATTERS

c. Present the Correlation Heatmap and explain two (2) insights – 2%

Correlated columns, 0.8 and above :

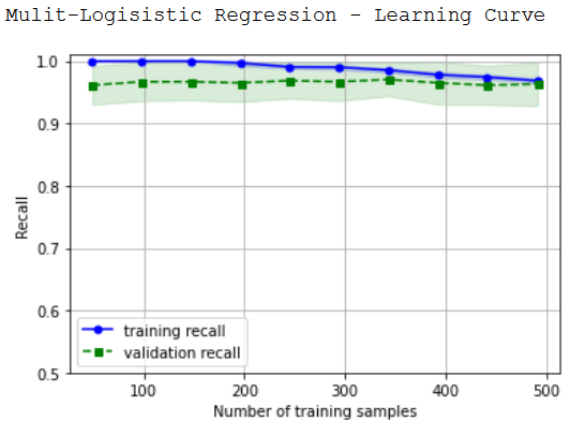
1. There is a strong relation between the class and UofCsize, UofCshape and barre Nuclei.
   - Class,UofCSize →.82
   - Class ,UofCShape → .82
   - Class ,Bare Nuclei →.82

2. The size and shape are strongly correlated, we can keep one of them in the analysis
   - UofCSize and UofCShape →.91

c. Present the Learning Curve for the Logistical Regression standard model and identify two (2) insights – 2%

1. There is high accuracy the model tells us that it could be an Overfitting model.
1. The number of features are so many(training and validation are so close to each other.

Average Bias: 0.02
Average Variance: 0.01



DURHAM COLLEGE
SUCCESS MATTERS

d. Present and Explain three (3) key insights from the classification report metrics (i.e. Precision, Recall, F1) for the Optimized Logistical Regression Model – 7%

Precision, out of all predictions , we got 98% correct.
Is about being precise.how accurate the model.
Recall , out of all correct guesses(truth) , we got 98% correct.
When predicts positive , how often it is correct.
F1, the harmonic mean of Precision and Recall, we got 98%.

e. Present and Explain two (2) key insights from ROC/AUC Curve (Optimized Model) – 2%

The ROC is plotted with the True Positive rate(y-access) against the FPR(False positive rate).

ROC is a curve of probability , ROC in our case is 1 , it is a perfect model ,it shows the trade off between sensitivity (True Positive Rate)and False Positive Rate(1-Specifity).Classifier that give curves closer to the top-left corner indicate better performance.
Same for the area under the graph AUC , which means that the two classes don't overlap and are well separated.This again supports our assumption of over fitting.

f. State and explain two (2) recommendations for Mr. John Hughes for next steps. – 2%

1. Collect more training data,
2. Reduce the complexity of the model by taking important features only.
3. Increase the regularization parameter for regularized model.

DURHAM COLLEGE
SUCCESS MATTERS