Symphony Hopkins

DSCI 502: R Programming

Professor Xiaobo Li

December 6, 2022

Project 8

1. Load the dataset in breast\_cancer\_data.csv. Call the loaded data breast\_cancer\_data. Make sure that you have the directory set to the correct location for the data.
   1. See R script.
2. Define a user defined function BoxplotPredictorOnTarget with two arguments, the target and one predictor to plot the box plot of predictor based on different category of the target. Then use this user defined function to generate the box plot:
   1. area\_mean against DiagnosisChart, box and whisker chart

      Description automatically generated
   2. area\_se against DiagnosisChart

      Description automatically generated
   3. texture\_mean against DiagnosisChart, box and whisker chart

      Description automatically generated
3. Build the following logistic models to forecast the Diagnosis and recommend the best model based on McFadden/pseudo R squared to the management.
   1. forecast Diagnosis using area\_meanGraphical user interface, text, email

      Description automatically generated
   2. forecast the Diagnosis using area\_mean and area\_seText

      Description automatically generated
   3. forecast the Diagnosis using area\_mean, area\_se and texture\_meanText

      Description automatically generated
   4. forecast the Diagnosis using area\_mean, area\_se, texture\_mean and concavity\_worstText

      Description automatically generated
   5. forecast the Diagnosis using area\_mean, area\_se, texture\_mean, concavity\_worst and concavity\_mean.Graphical user interface, text

      Description automatically generated
   6. Based on the McFadden/pseudo r-square values, the best logistic model is 3E with a score of 0.79.