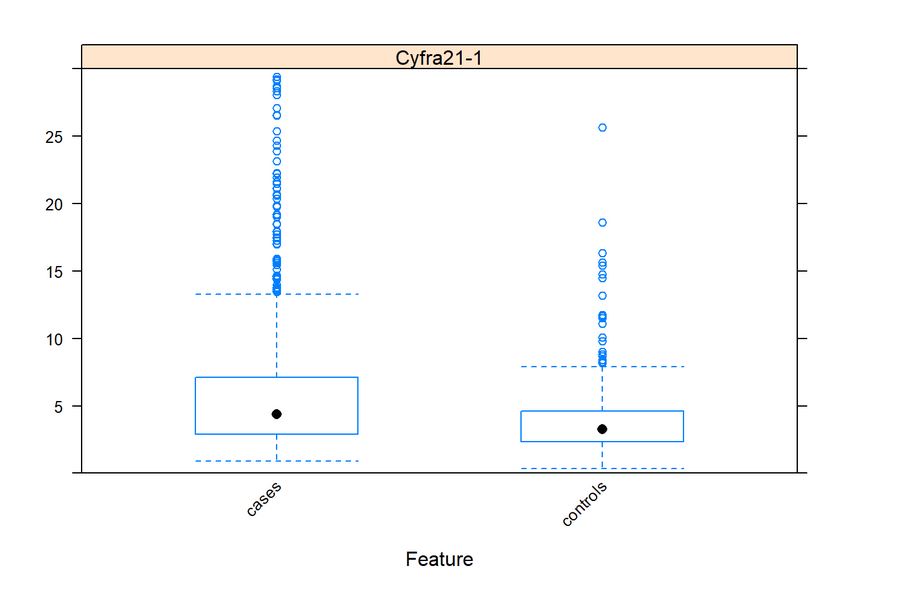
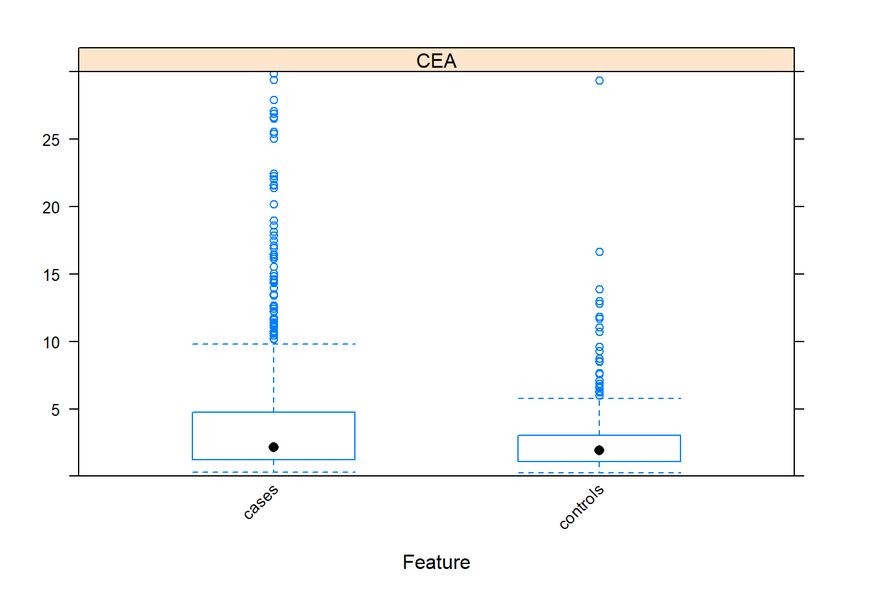
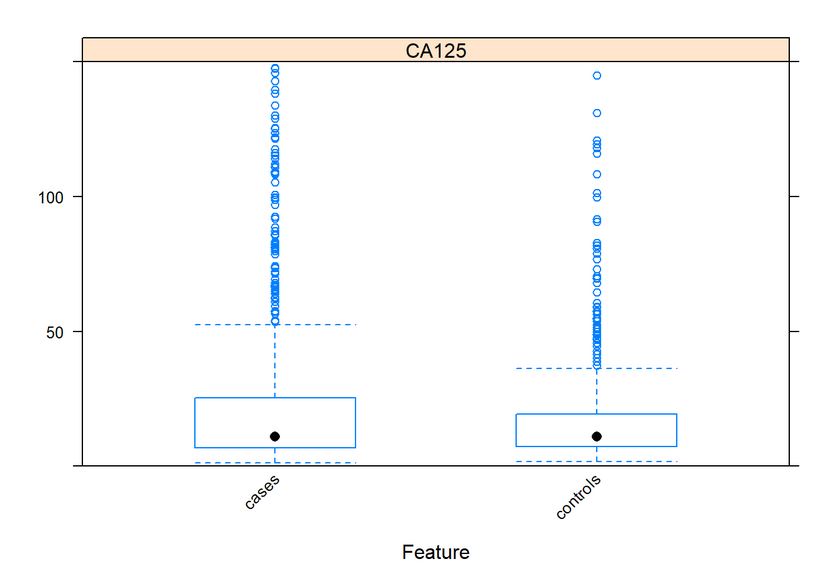
**Lung Cancer Data Analysis**

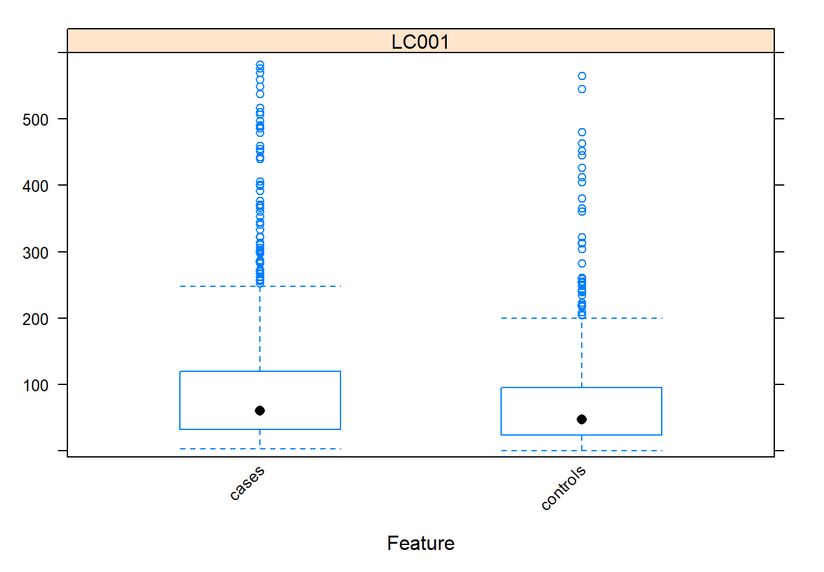
Date: March 31, 2022

* Data:
  + 5 columns and 1366 rows
  + No missing values
  + Columns:
    - Class : 5 classes
      * advanced lung cancer - 384 rows
      * Benign lung tumor - 215 rows
      * Early lung cancer - 464 rows
      * Other cancers - 92 rows
      * Pneumonia -211 rows
    - Cyfra21-1 : CYRFA 21-1 (cytokeratin 19 fragment) has been demonstrated as clinically useful in the prognostication and monitoring of non-small cell lung cancer (NSCLC).
    - CEA : The carcinoembryonic antigen (CEA) test measures the level of CEA in the blood.
    - CA125 : A CA 125 test measures the amount of the protein CA 125 (cancer antigen 125) in the blood.
    - LC001: Level of protein in the serum.
* Data analysis:
  + Feature box plot was plotted for cases and controls:



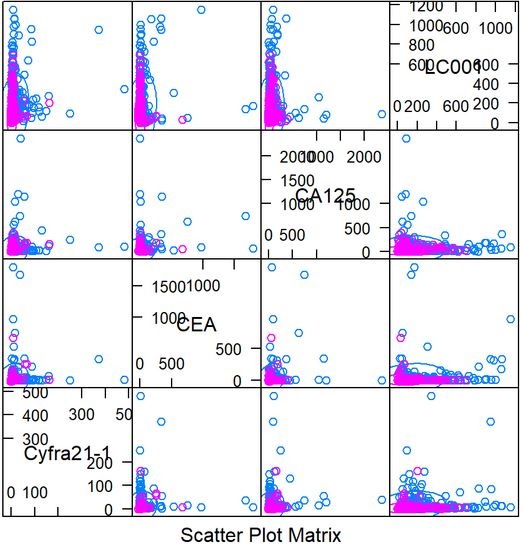






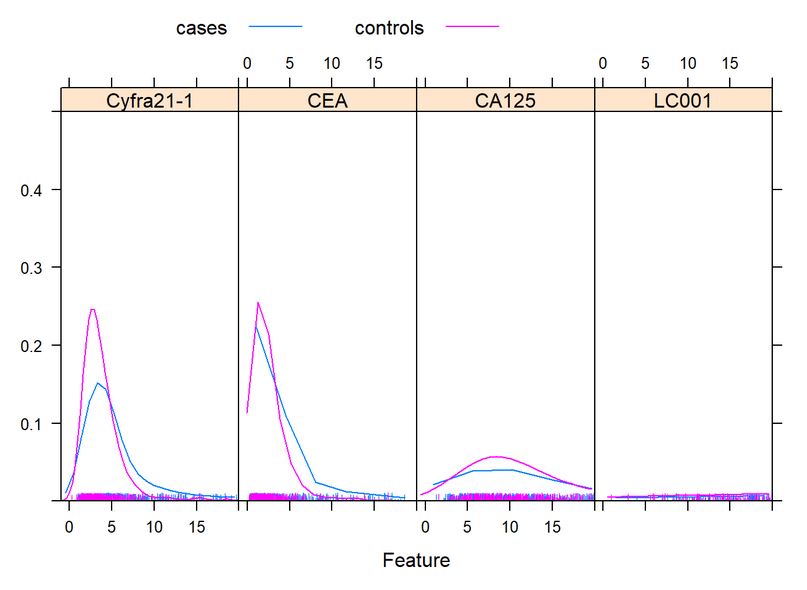


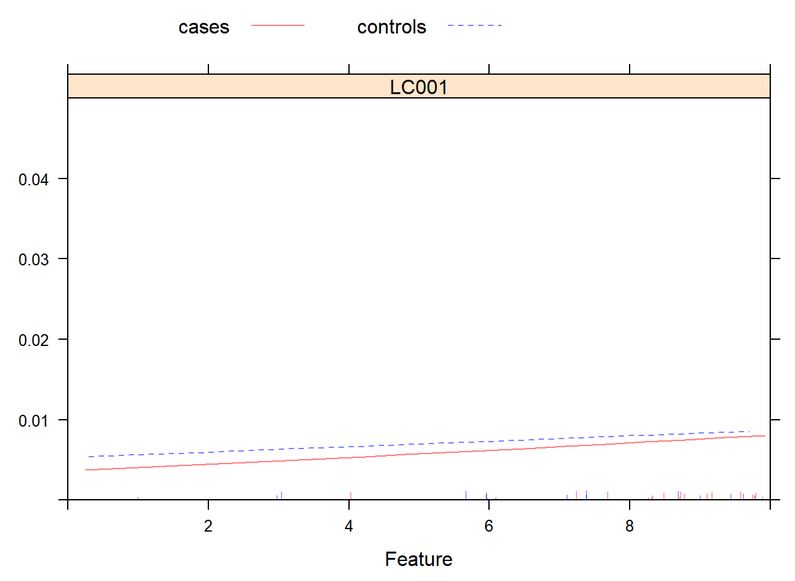
* + Scatter plot matrix:





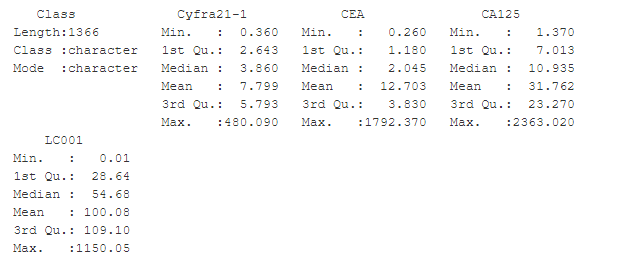
* + Density plot is also plotted for data



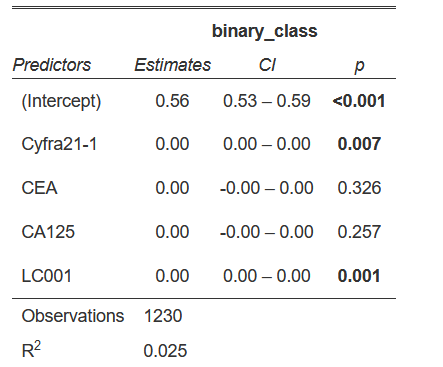




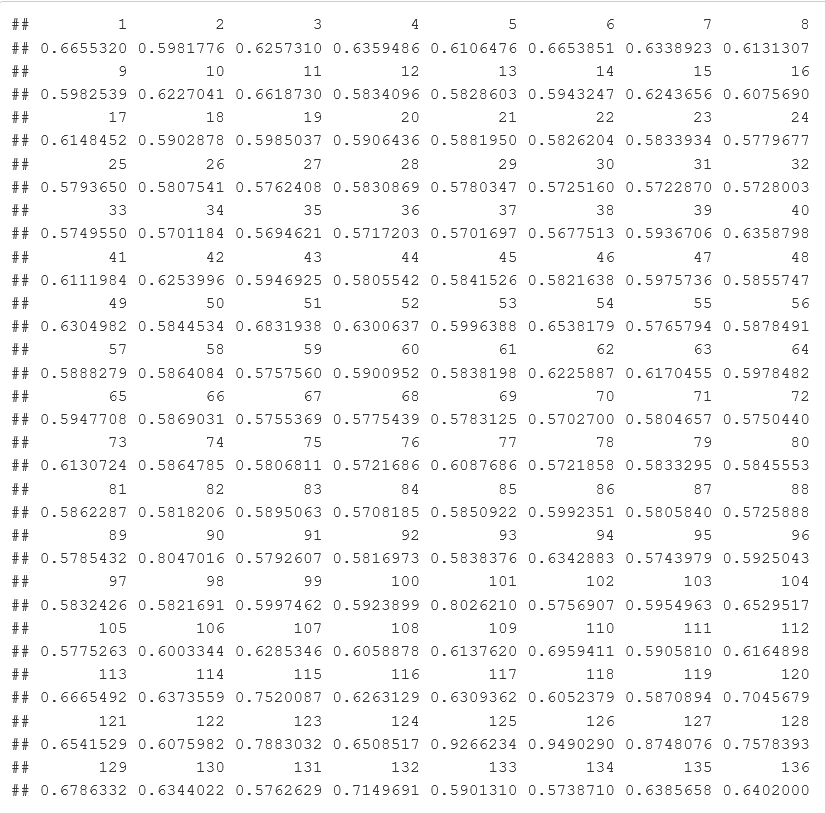
* + Summary statistics



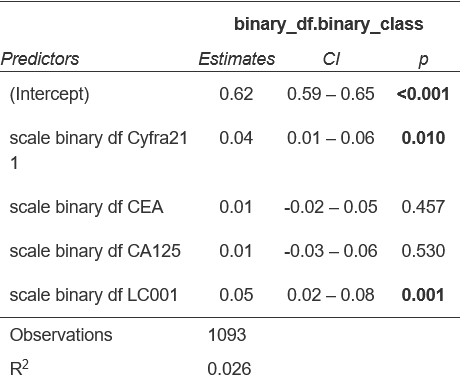
* Models:
  + Multinomial regression is used as it is a multiclass problem.
  + In statistics, multinomial logistic regression is a [classification](https://en.wikipedia.org/wiki/Statistical_classification) method that generalises logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.A log-linear model is fitted, with coefficients zero for the first class (here **advanced lung cancer**).
  + Model was made on unscaled and scaled data and the results are demonstrated here.
    - Unscaled model
      * Class is to be predicted with Cyfra21-1, CEA CA125 and LC001 as the independent variables.

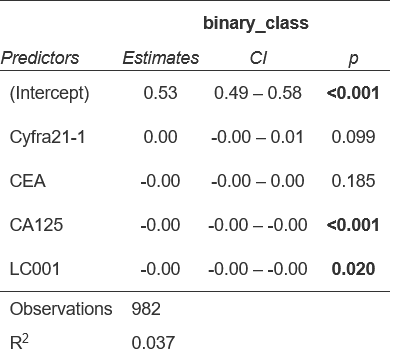


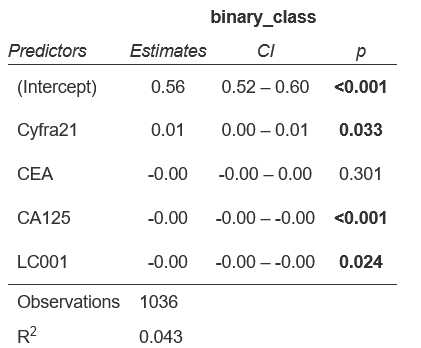
* Unscale model



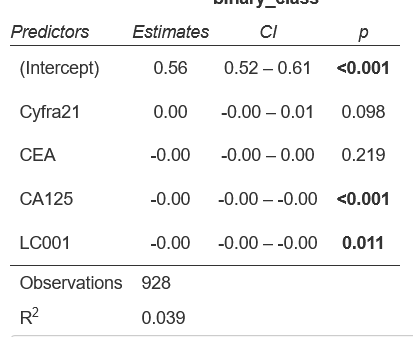
* predicted model 1 response



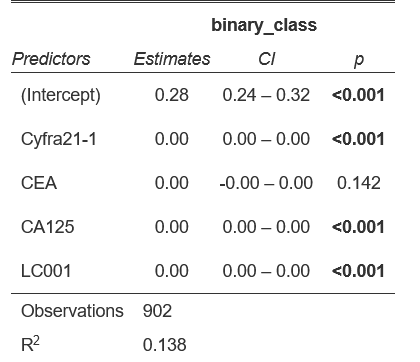
* Scaled model
* Case 1 model results



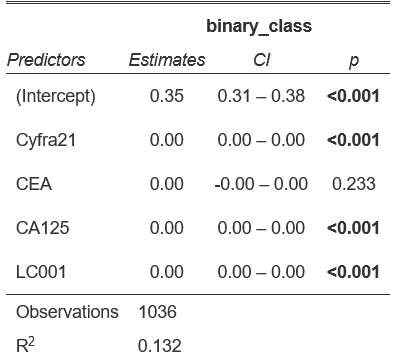
* Case 1 over sampling model results



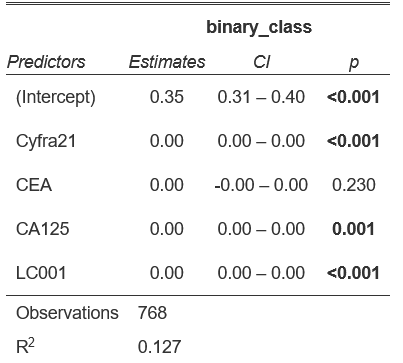
* Case 1 under sampling model results



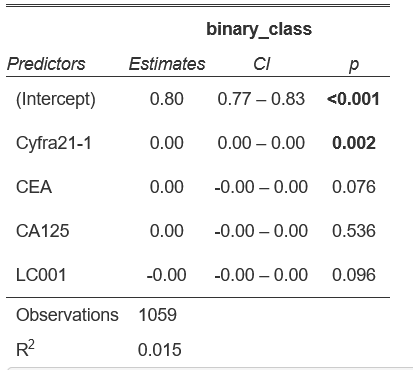
Case 2 model results



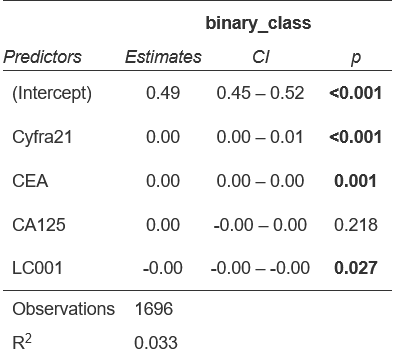
* Case 2 over sampling data results



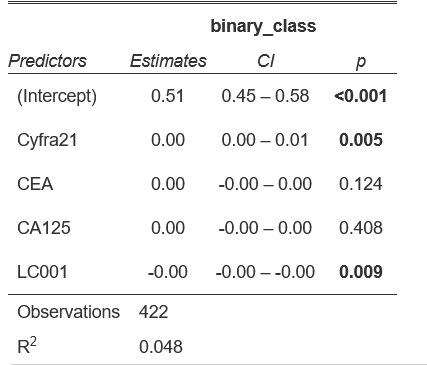
Case 2 under sampling data results



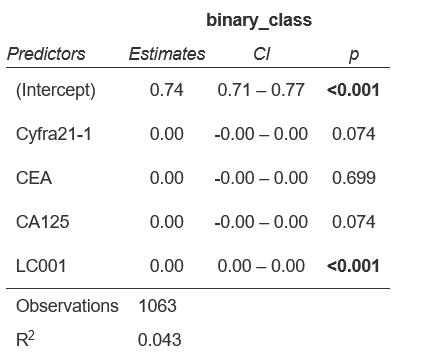
* Case 3 model results



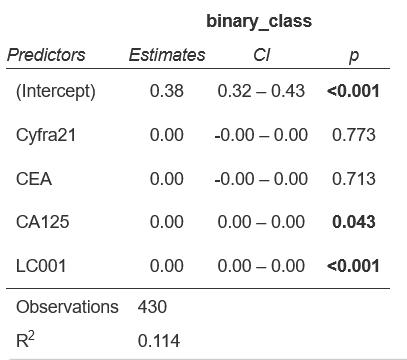
* Case 3 over sampling model results



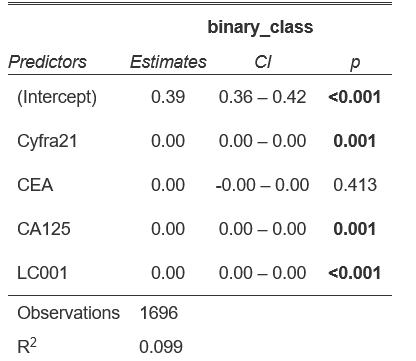
* Case 3 under sampling model results



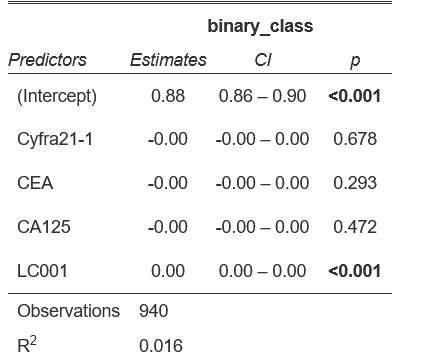
* Case 4 model results



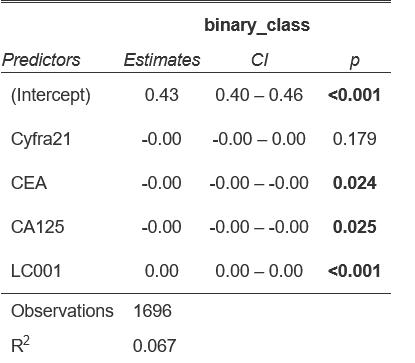
* Case 4 under sampling model results



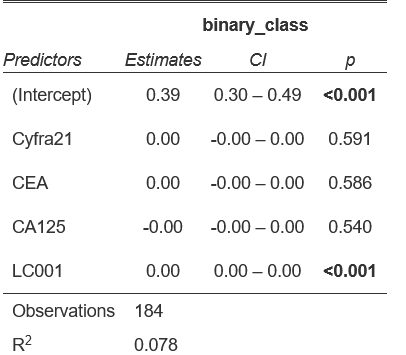
* Case 4 over sampling model dataset



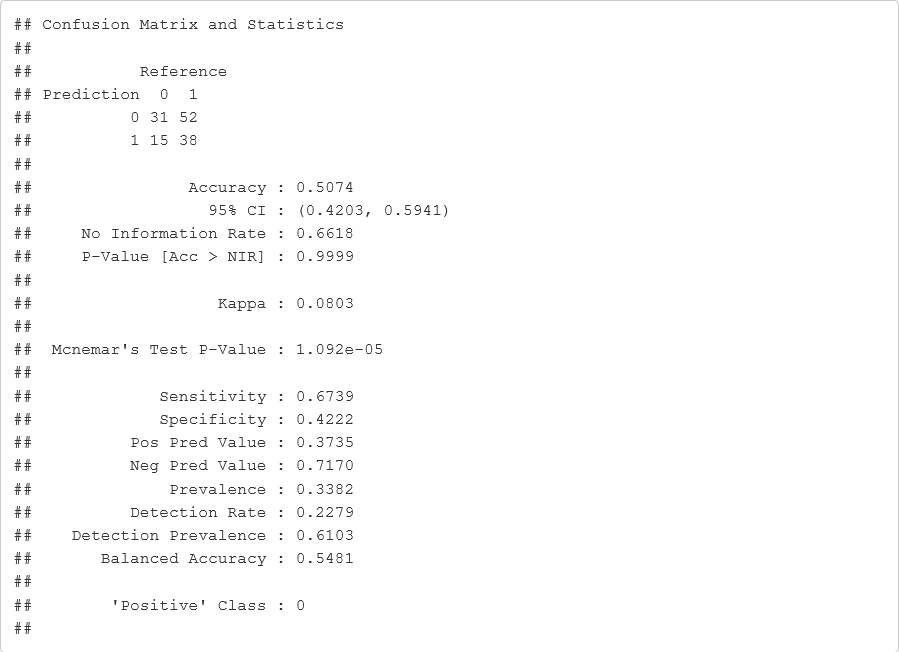
* Case 5 model results



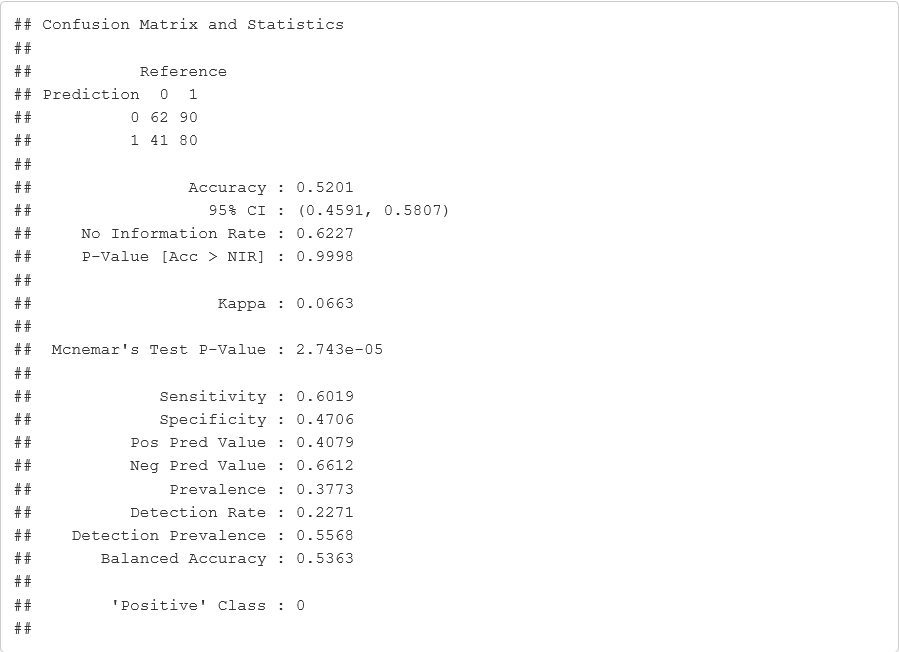
* Case 5 over sampling model results



* Case 5 under sampling model results
  + Model was trained on 80 percent of the data and the rest 20 percent was used to predict and the results are as follows.
    - These are the class labels.
      * cases - 0
      * Control - 1
    - Confusion matrix and other stats are shown below:



* + - * Confusion matrix of unscaled model



Confusion matrix of scaled model

* + Considering as 5 class, and considering the imbalance data set, data seems to be reliable.