STAT469 assignment #3

name : Bill Co

I. Overview

This Assignment is the extension of the previous one, in which I will introduce two more models, namely Multi-task Prediction(MTPS) and Random Forest. Furthermore, I have chosen

II. Data preparation and Model assessment

a. Resampling

To ensure data stability, I have decided to use resampling to generate simulation data for model fittings. After some testing, I can make sure this method is identical to stratification, but some observations are repeated.

b. Simulations

There are going to be 50 simulations, each with 5 fold Cross Validation.

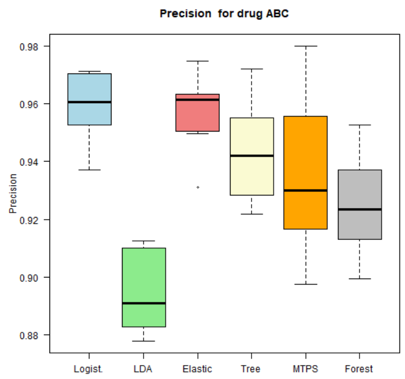
c. Metrics

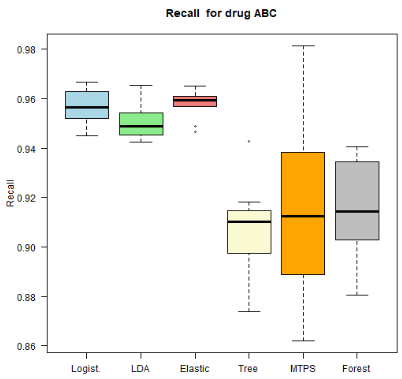
For productivity, some matrices are used to store metrics. Since the dataset shows class imbalance, metrics like Accuracy and Misclassification rate can be misleading. Thus, I excluded them from the assessment, keeping only Precision, Recall and F1 score.

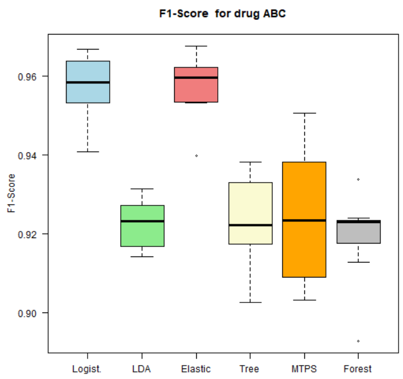
III. Result discussion (with visualization)

1. Individual Drug

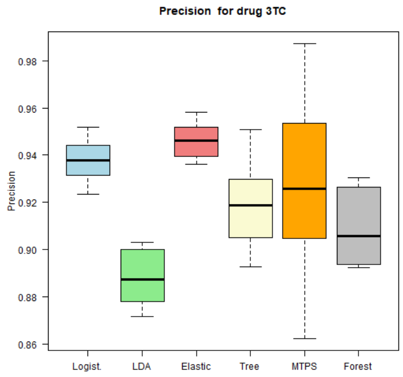
a. ABC

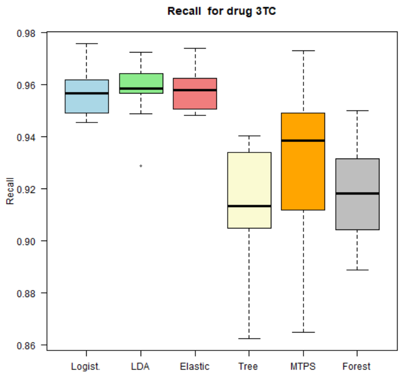


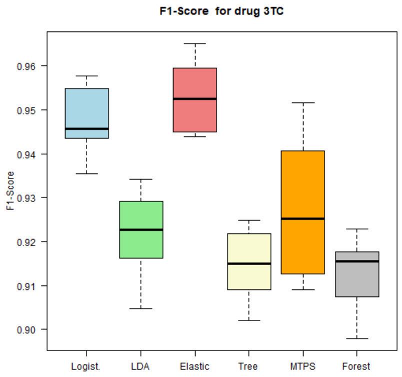




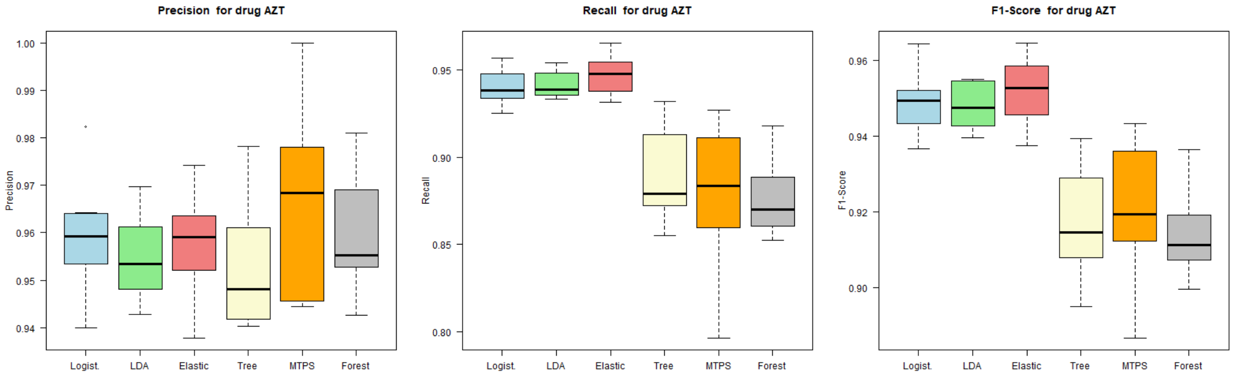
b. 3TC



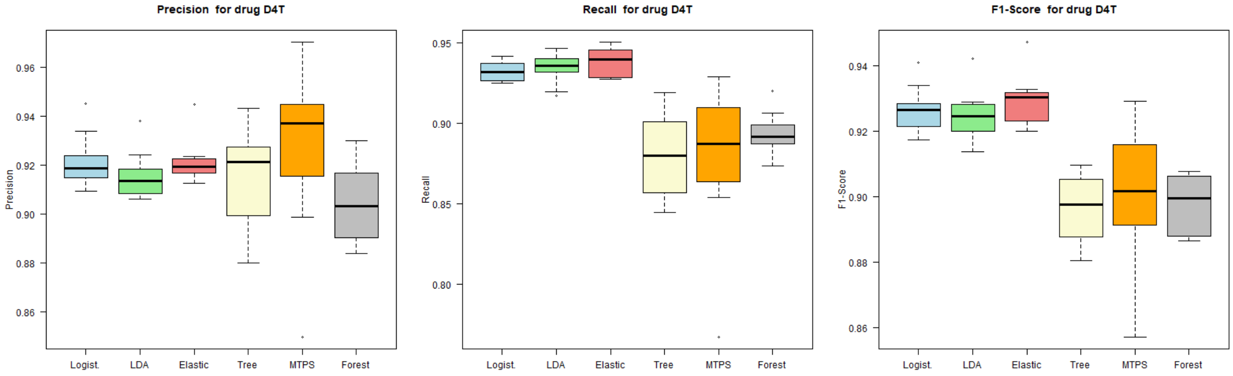




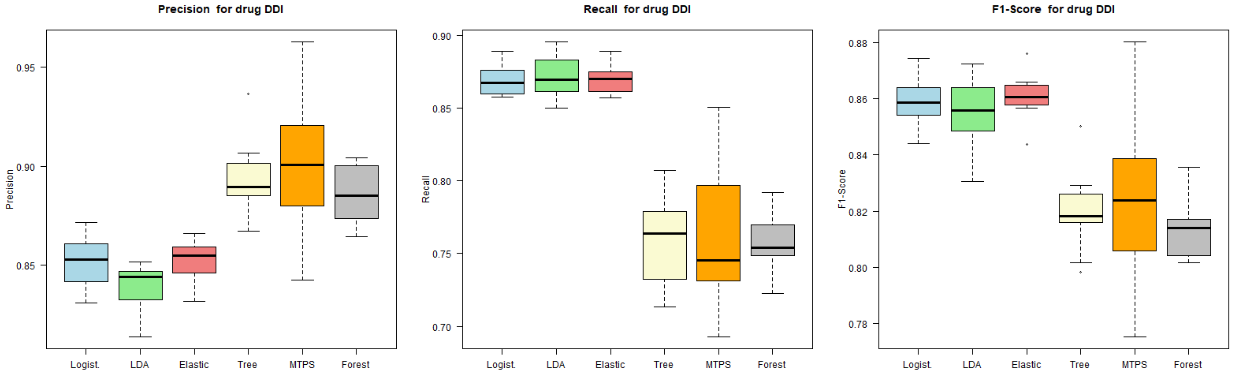
c. AZT


Precision for drug AZT
Precision
1.00
0.99
0.98
0.97
0.96
0.95
0.94
Recall for drug AZT
Recall
0.95
0.90
0.85
0.80
F1-Score for drug AZT
F1-Score
0.96
0.94
0.92
0.90
Logist.
LDA
Elastic
Tree
MTPS
Forest

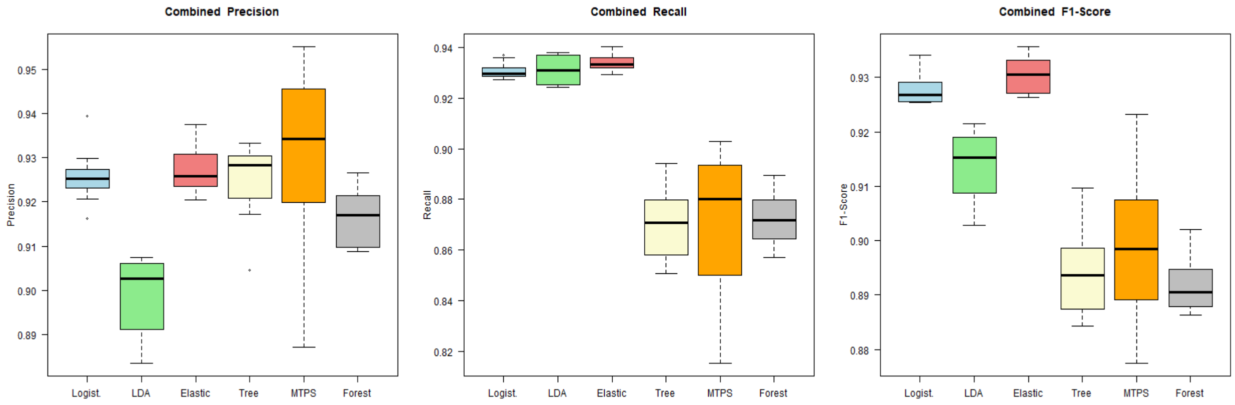
d. D4T


Precision for drug D4T
Precision
0.96
0.94
0.92
0.90
0.88
0.86
Recall for drug D4T
Recall
0.95
0.90
0.85
0.80
F1-Score for drug D4T
F1-Score
0.94
0.92
0.90
0.88
0.86
Logist.
LDA
Elastic
Tree
MTPS
Forest

e. DDI


Precision for drug DDI
Precision
0.95
0.90
0.85
Recall for drug DDI
Recall
0.90
0.85
0.80
0.75
0.70
F1-Score for drug DDI
F1-Score
0.88
0.86
0.84
0.82
0.80
0.78
Logist.
LDA
Elastic
Tree
MTPS
Forest

2. Combined Drug



The above visualizations are the assessments grouped by Drugs. As shown, Elastic Net and Logistic Regression are very stable across all measurements.

While MTPS shows a very high precision, the Recall and F1-Score are overkilled. LDA has a quite reasonable Recall but performs not very well in the Precision and F1-Score.

Tree and Forest are not good candidates for the best model in this scenario.

3. Hypothesis Testing

After conducting hypothesis, I concluded that:

- At the probability of 23%, there is no clear difference between the 2 models in Precision. However, Elastic Net performs better overall.

Thus, combining 2 criteria above, I decided **Elastic Nets should be used for predicting Drug resistance for HIV disease.**