STAT469 Assignment #2

name : Bill Co

I. Overview

This Assignment is the extension of the previous one, in which we will introduce two more models, namely Classification trees and Elastic Nets. Additionally, we are not including KNN this time.

After taking into consideration, I have observed 3 things about Assignment#1:

1. The code was unnecessarily lengthy.
2. The method was partially correct, but something was still missing.
3. I eyeballed the Model assessments to curate the "best" model, which is not a good idea. Hypothesis testing and Visualizations should have been used for model selection.

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.

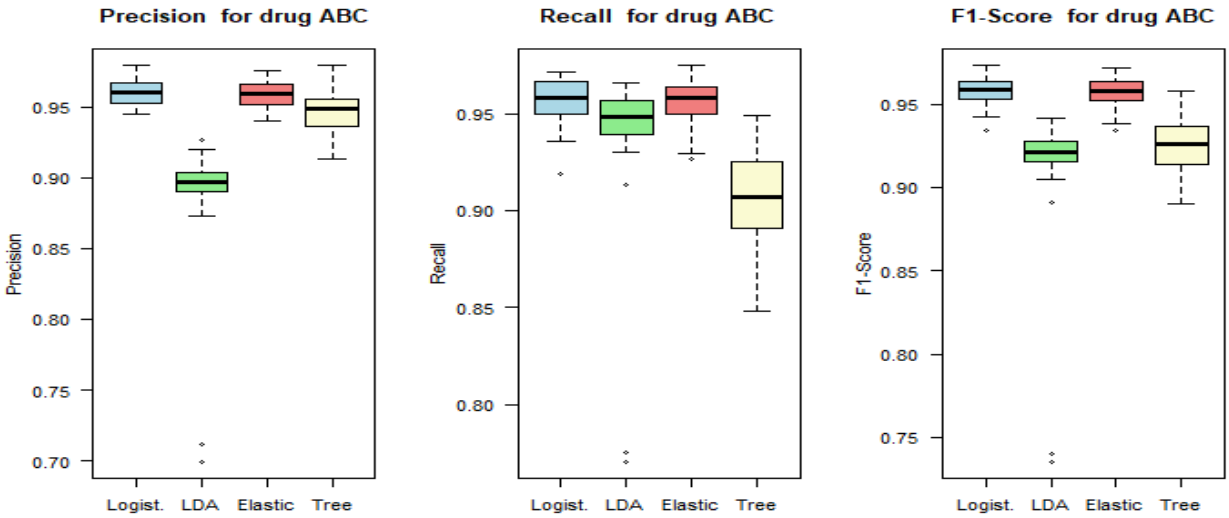
c. Metrics

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.

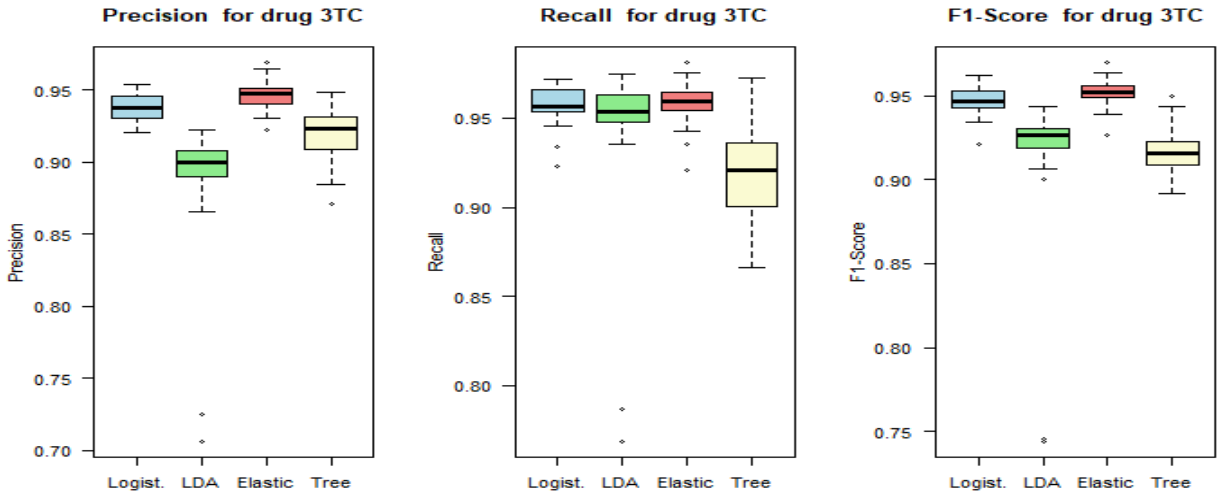
III. Result discussion (with visualization)

1. Individual Drug

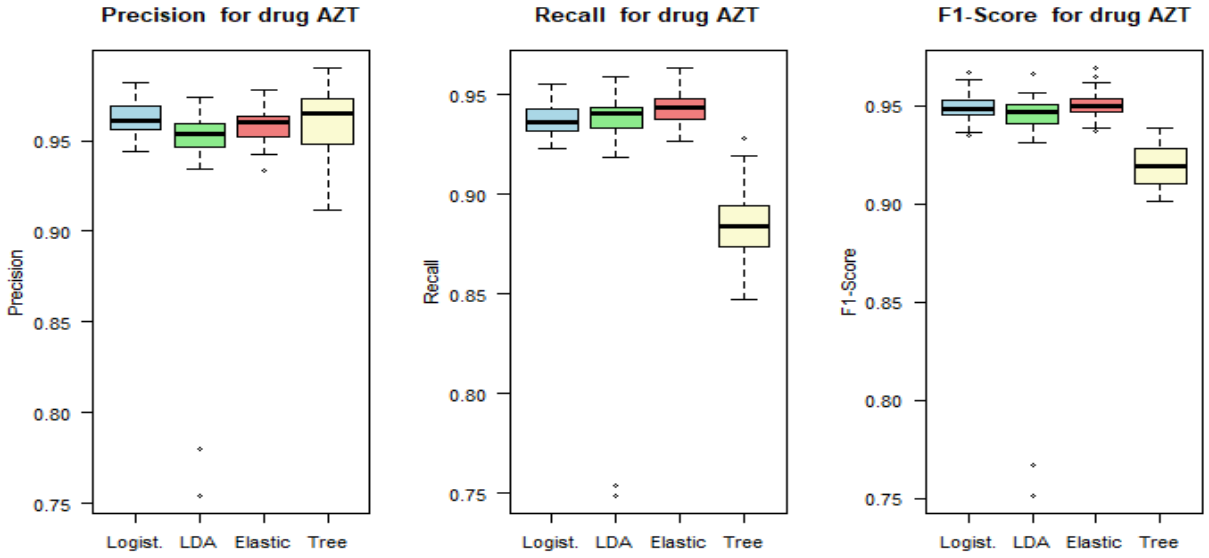
a. ABC



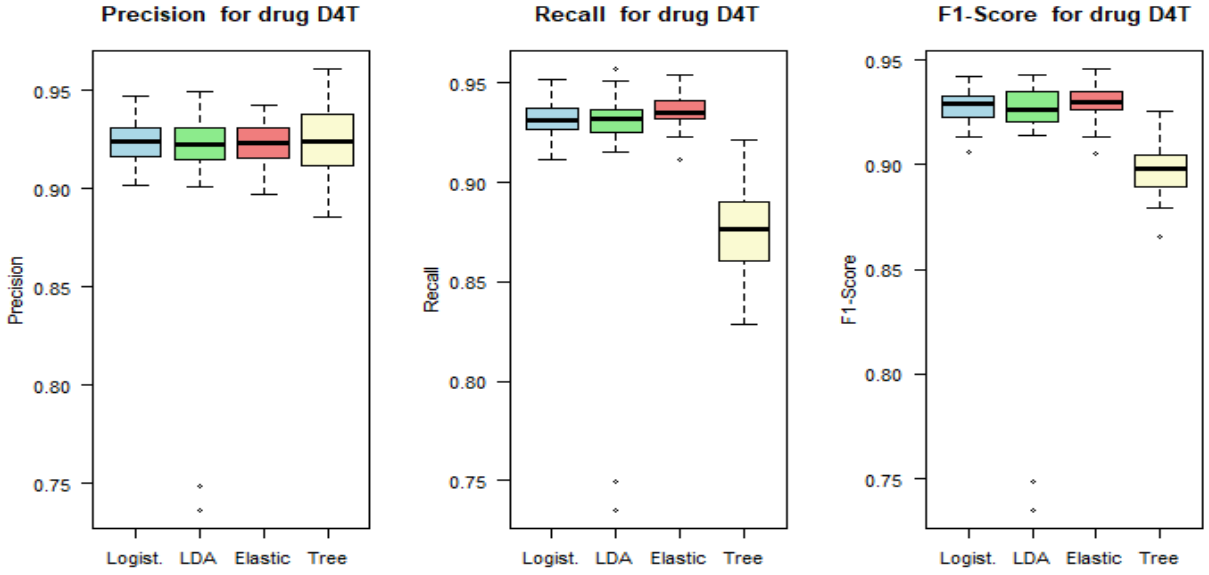
b. 3TC



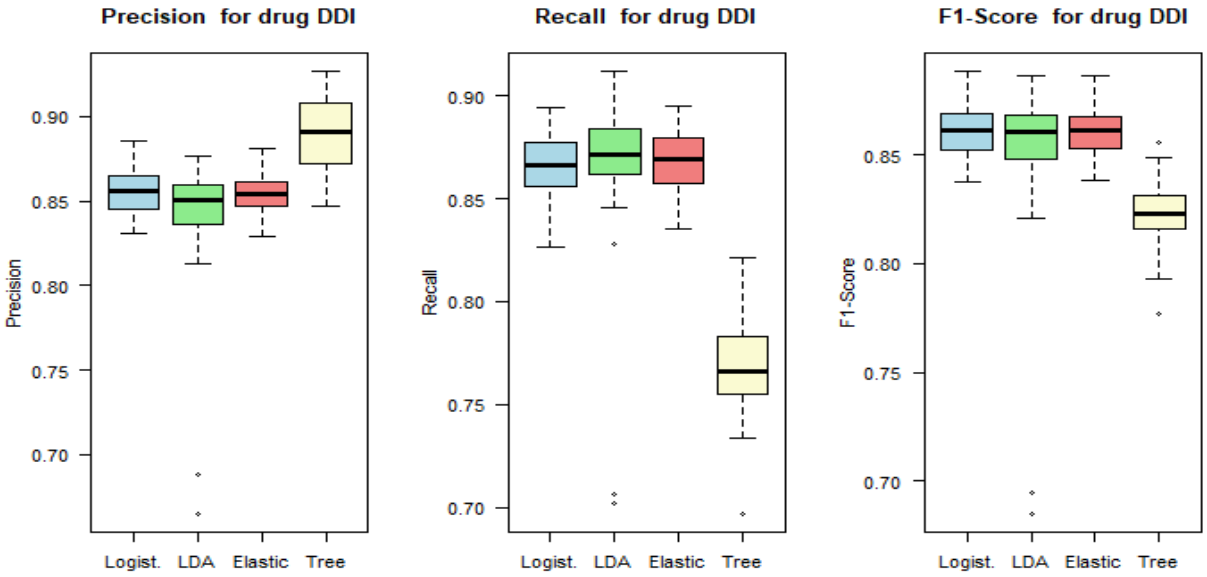
c. AZT


Precision for drug AZT
Precision
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree
Recall for drug AZT
Recall
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree
F1-Score for drug AZT
F1-Score
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree

d. D4T


Precision for drug D4T
Precision
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree
Recall for drug D4T
Recall
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree
F1-Score for drug D4T
F1-Score
0.95
0.90
0.85
0.80
0.75
Logist.
LDA
Elastic
Tree

e. DDI

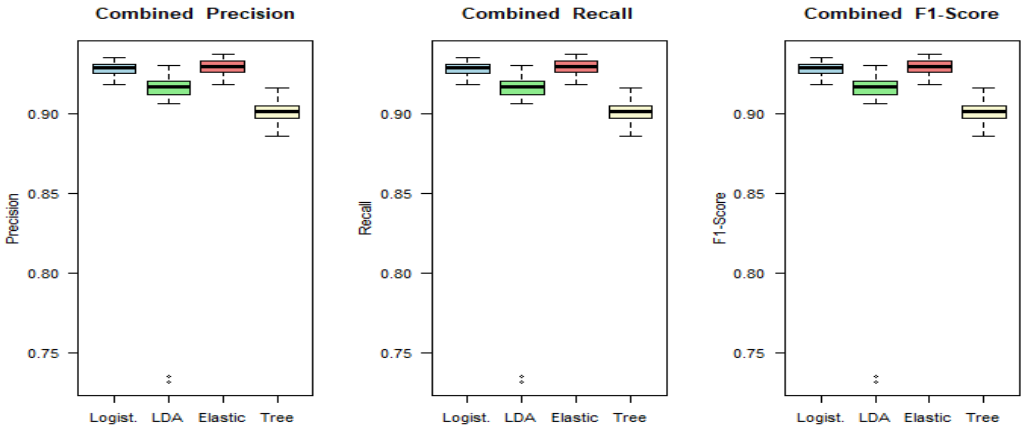


Overview:

- For every drug, trees outperform only once in the precision, plus they show a lot of weaknesses compared to other models so I chose to not consider Classification trees anymore.

- Other models show some overlaps, so we will group all drugs together for a simpler explanation.

2. Combined Drug



The above visualizations are the assessments grouped by Drugs. As shown, Trees perform not very well in total. LDA is not exceptional, while in individual cases, it shows some overlaps with Logistic (L1) regression and Elastic Nets, but group by demonstrates that it should not be a candidate for the final model.

We will conduct hypothesis testing for F1-Score on LDA with Classification Trees and Elastic Nets.

3. Hypothesis Testing

After conducting hypothesis testing, the results show:

- Median F1 score of LDA is 0.917, which is less than that of Elastic Nets (0.93) but greater than Classification Trees (0.897).
- Null Hypotheses (True location shift is 0) are rejected in both cases, which suggest that Elastic Nets perform the best compared to the other 2 models.

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