CASE STUDY

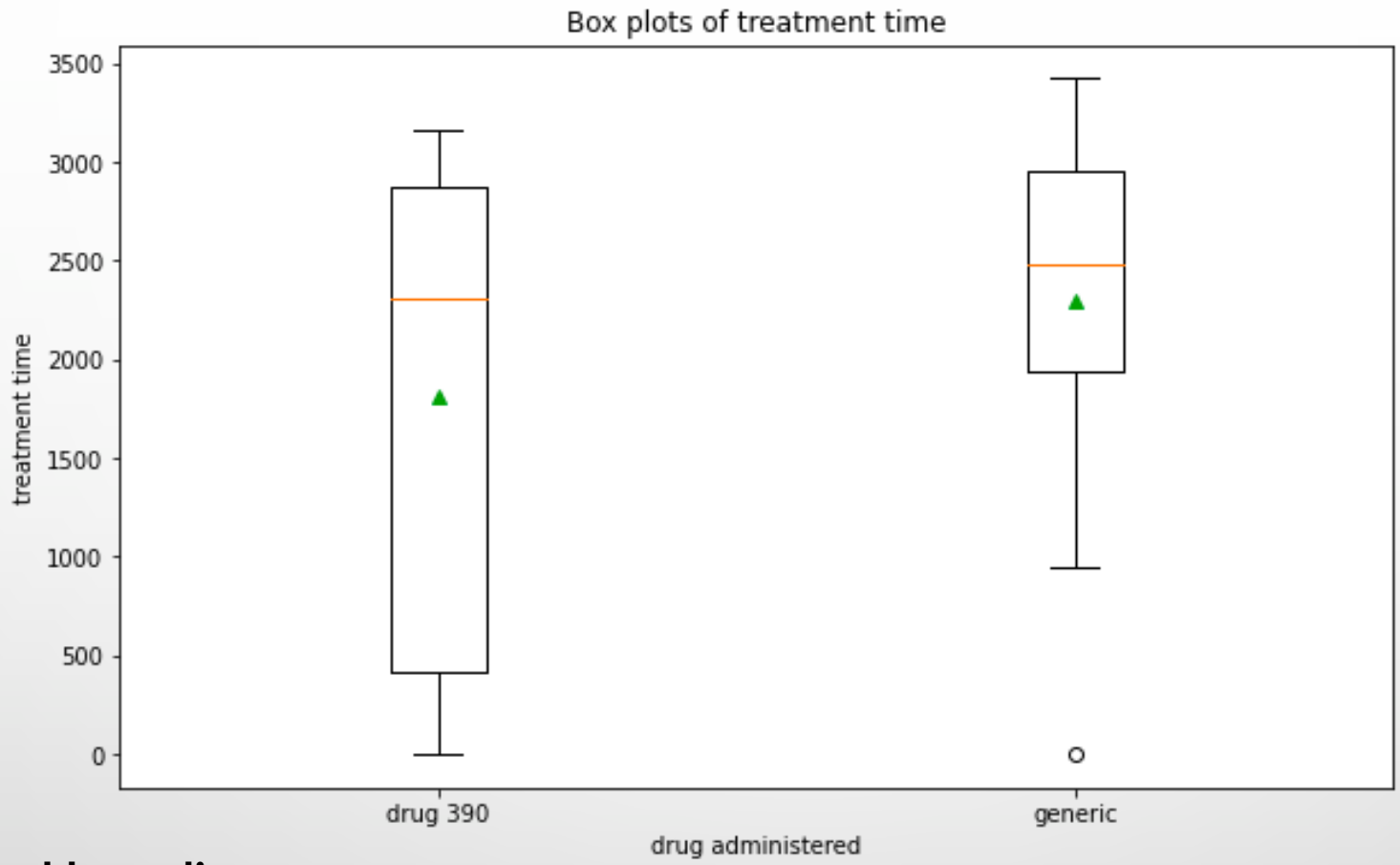
DRUG TREATMENT TIME, AND HEART DISEASE CAUSATIVE FACTOR PREDICTION

QUESTION 1

A. THE LENGTH OF THE TREATMENT CAN BE COMPUTED FROM THE DRUG ADMIN DATE SHEET, WE GROUP BY THE PATIENT ID, AND SUBTRACT THE FIRST DATE HE WAS ADMINISTERED THE DRUG FROM THE LAST DATE.

NOTE: FOR 3 PATIENTS (IDS: 2634, 6837, AND 6922) WE HAD ONLY ONE DATE, AND BY DEFAULT THEIR TREATMENT TIME IS TAKEN AS 0.

QUESTION 1 (B)

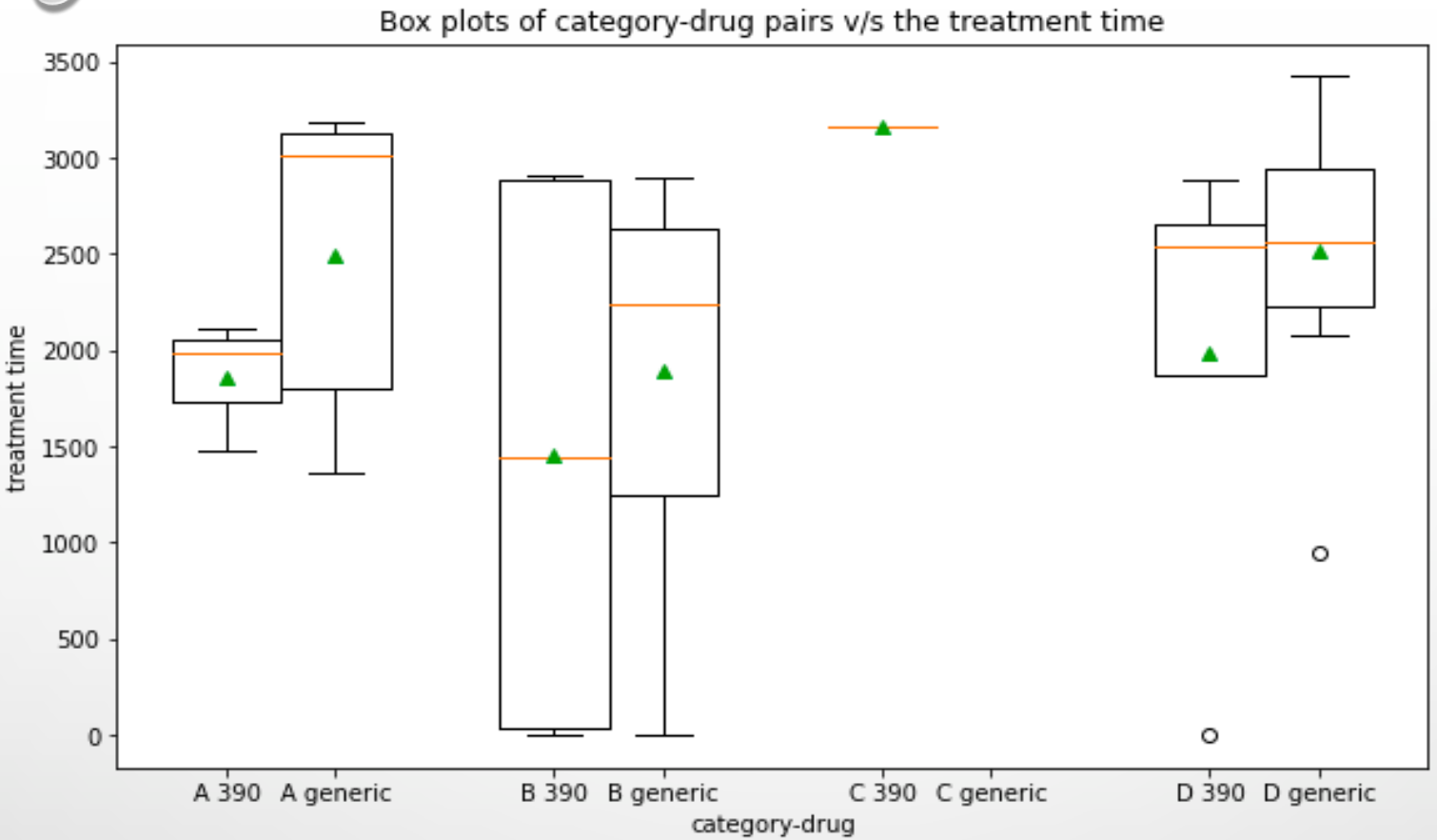


Green triangle gives the mean

- **Comparable medians**
- **Low mean for drug 390**
- **Drug 390 is more spread between the first quartile and the median**
- **The extremums are less for the drug 390**

QUESTION 1 (C)

*Appendix 1 for category definitions



- **The means and the medians are higher for generic drug (longer treatment time)**
- **In category B the treatment time is symmetric, and the extremums are same for both drugs.**
- **There is only 1 patient of category C in the dataset.**
- **Only 4 patients of category D are administered drug 390, smallest point is inside the lower quartile.**

QUESTION 1(D)

- A good target to predict from this dataset is treatment time.
- For a new patient, we can first identify his HER2 status, ER status, and PR status.
- Make two test points, one with drug 390 flagged as 1, other with 0.
- See which drug predicts the smallest treatment time, and use it.
- Doing feature augmentation can help in prediction (age, gender etc.)

QUESTION 2 (A)

MODELS USED

Logistic Regression

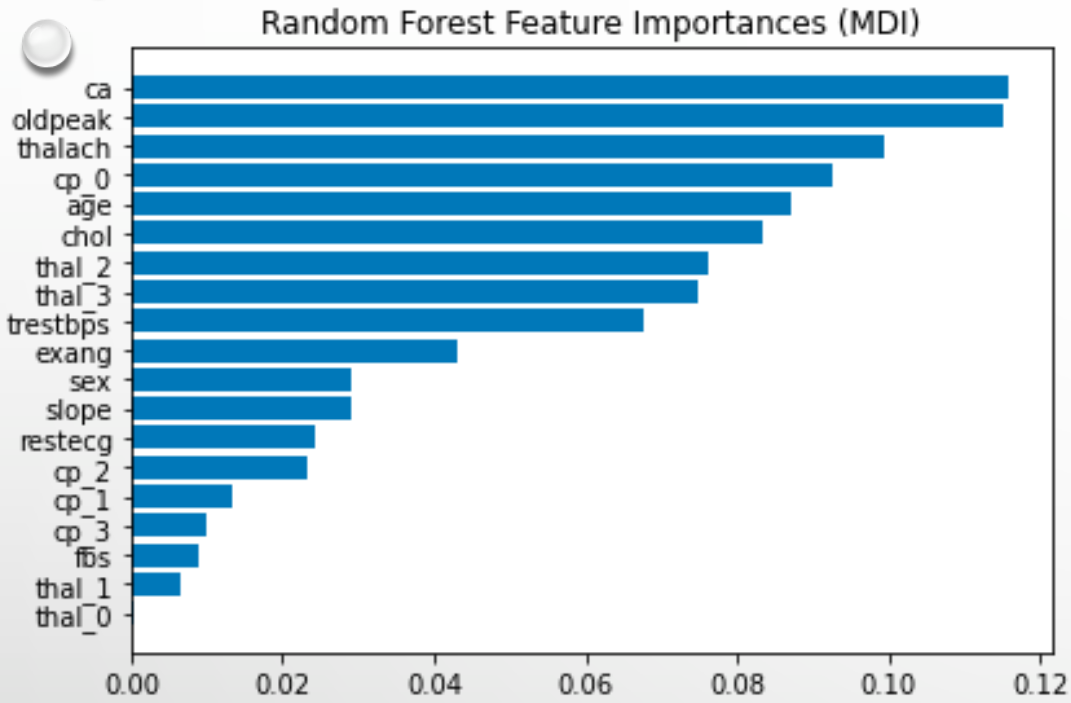
- Faster to train.
- Easy interpretability.
- Coefficients can be used to identify feature importance. (continuous features are scaled)

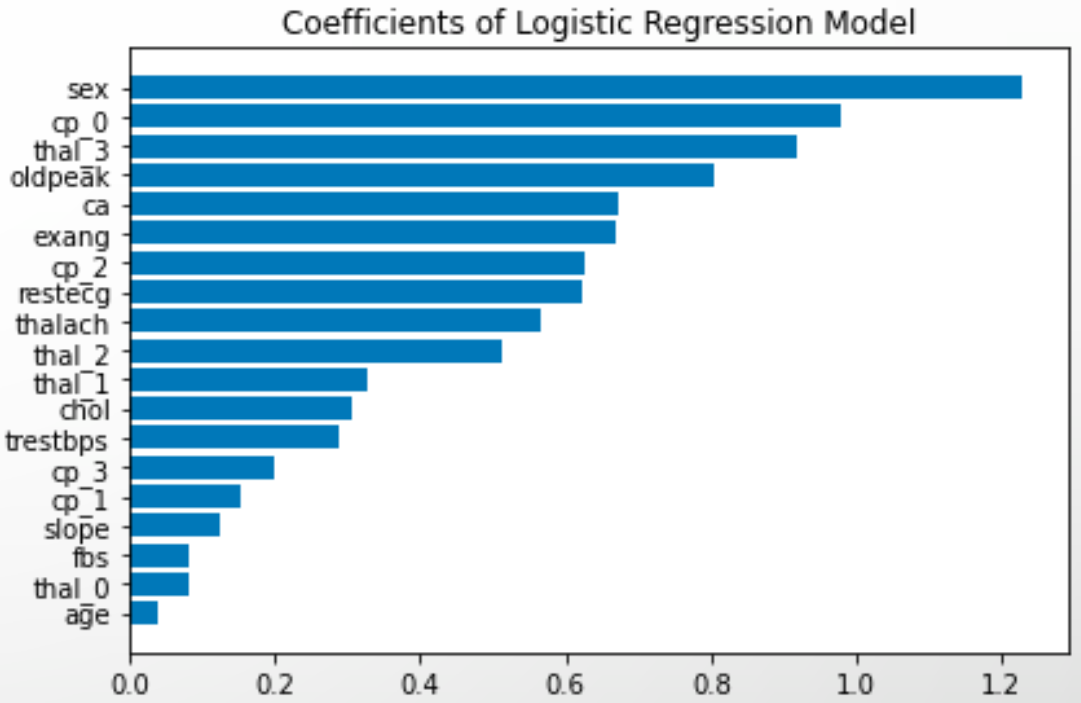
Random Forest

- Theoretically can achieve low bias and low variance.
- No scaling or encoding required for categorical variables.
- We can get the feature importance from MDI (Mean Decrease in Impurity)

Note:
1. Accuracy was used as a metric as the data is nearly balanced.
2. For stepwise feature selection based on AIC see appendix.

QUESTION 2(B)

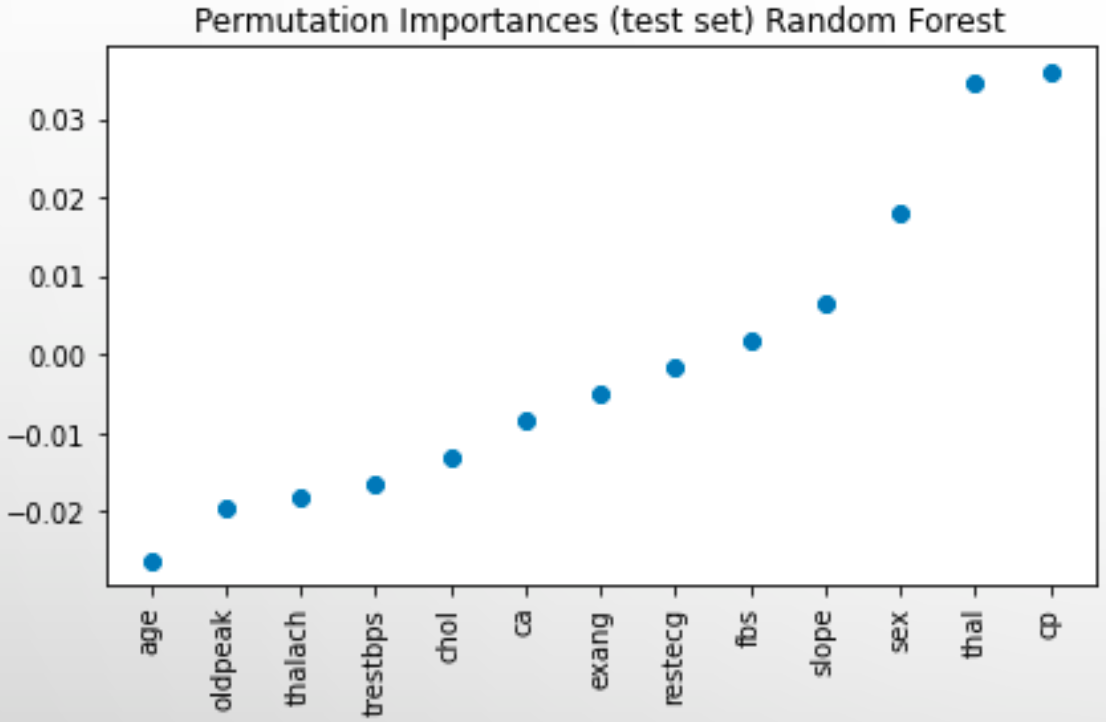


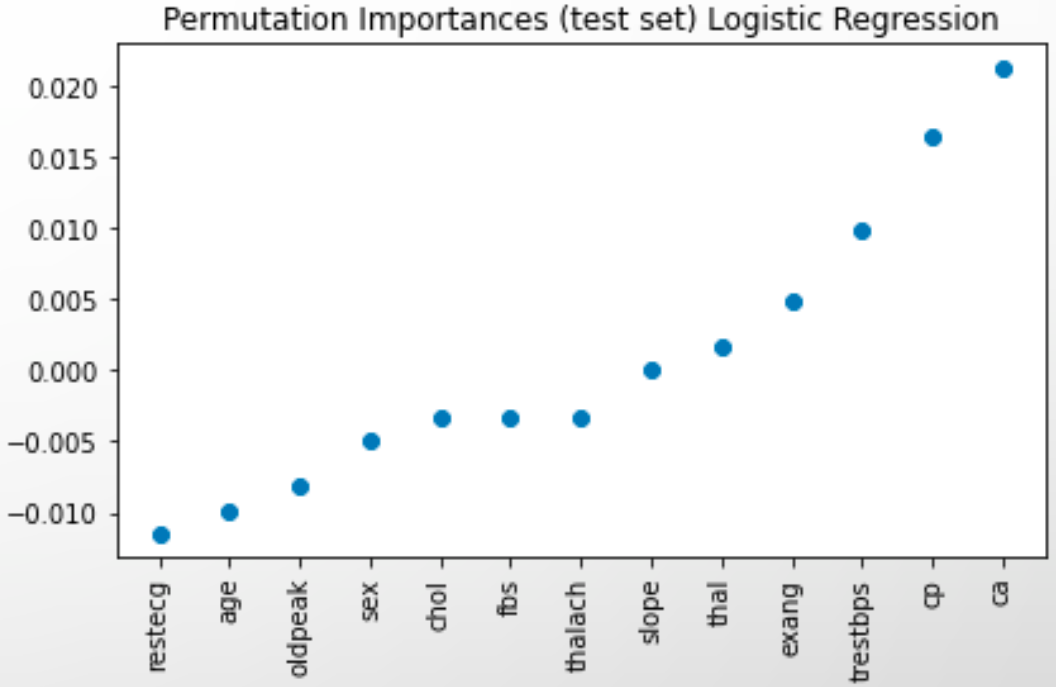


The features that are an intersection of the top 10 features from both the models are:
ca (no of major vessels), oldpeak (ST depression induced by exercise), thalach (maximum heart rate achieved), cp_0 (chest pain type zero), thal_2, thal_3, and exang (exercise induced angina).

Note: the cp and thal features were one hot encoded

PERMUTATION FEATURE IMPORTANCE ON TEST SET





The features that are an intersection of the top 8 features from both the models are:
ca, cp, exang, fbs, slope, and thal.

Note: here the features are not encoded, because first the feature is permuted then the encoding is done.

Significance Values and the coefficients of Logistic Regression from R model fit

```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5849  -0.3872   0.1551   0.5863   2.6249

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.450472   2.571479   1.342 0.179653
age         -0.004908   0.023175  -0.212 0.832266
sex         -1.758181   0.468774  -3.751 0.000176 ***
cp           0.859851   0.185397   4.638 3.52e-06 ***
trestbps    -0.019477   0.010339  -1.884 0.059582 .
chol        -0.004630   0.003782  -1.224 0.220873
fbs          0.034888   0.529465   0.066 0.947464
restecg      0.466282   0.348269   1.339 0.180618
thalach      0.023211   0.010460   2.219 0.026485 *
exang       -0.979981   0.409784  -2.391 0.016782 *
oldpeak     -0.540274   0.213849  -2.526 0.011523 *
slope        0.579288   0.349807   1.656 0.097717 .
ca          -0.773349   0.190885  -4.051 5.09e-05 ***
thal        -0.900432   0.290098  -3.104 0.001910 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 211.44  on 289  degrees of freedom
AIC: 239.44
```

Note: We fit generalized linear model with logit link (logistic regression) in R to obtain the p-values of the coefficients. (null: the coefficient =0). We did this, as the coefficients are just a crude way to get the feature importance. This tells us that **sex, cp, thalach, exang, oldpeak, ca, and thal** are significant at 5% significance level.

CONCLUSION

By taking an intersection of the top 6 features from the permutation feature importance and the significant
Coefficients from the GLM model, we find that **thal (thalassemia), ca(no of major vessel), cp (chest pain), and exang (exercise induced angina)** are main contributing factors towards heart disease in this dataset.

Please refer https://github.com/apalmk/pfizer_task/blob/master/Pfizer_task.ipynb for detailed analysis.

APPENDIX

APPENDIX 1

How I defined each category

HER2_status	HR_status	Category
0	0	*A*
0	1	*B*
1	0	*C*
1	1	*D*

STEPWISE FEATURE SELECTION BASED ON **AIC**

```
Initial Model:
target ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach +
    exang + oldpeak + slope + ca + thal

Final Model:
target ~ sex + cp + trestbps + restecg + thalach + exang + oldpeak +
    slope + ca + thal
```