TALHA MUHAMMAD

HEART DISEASE ASSIGNMENT

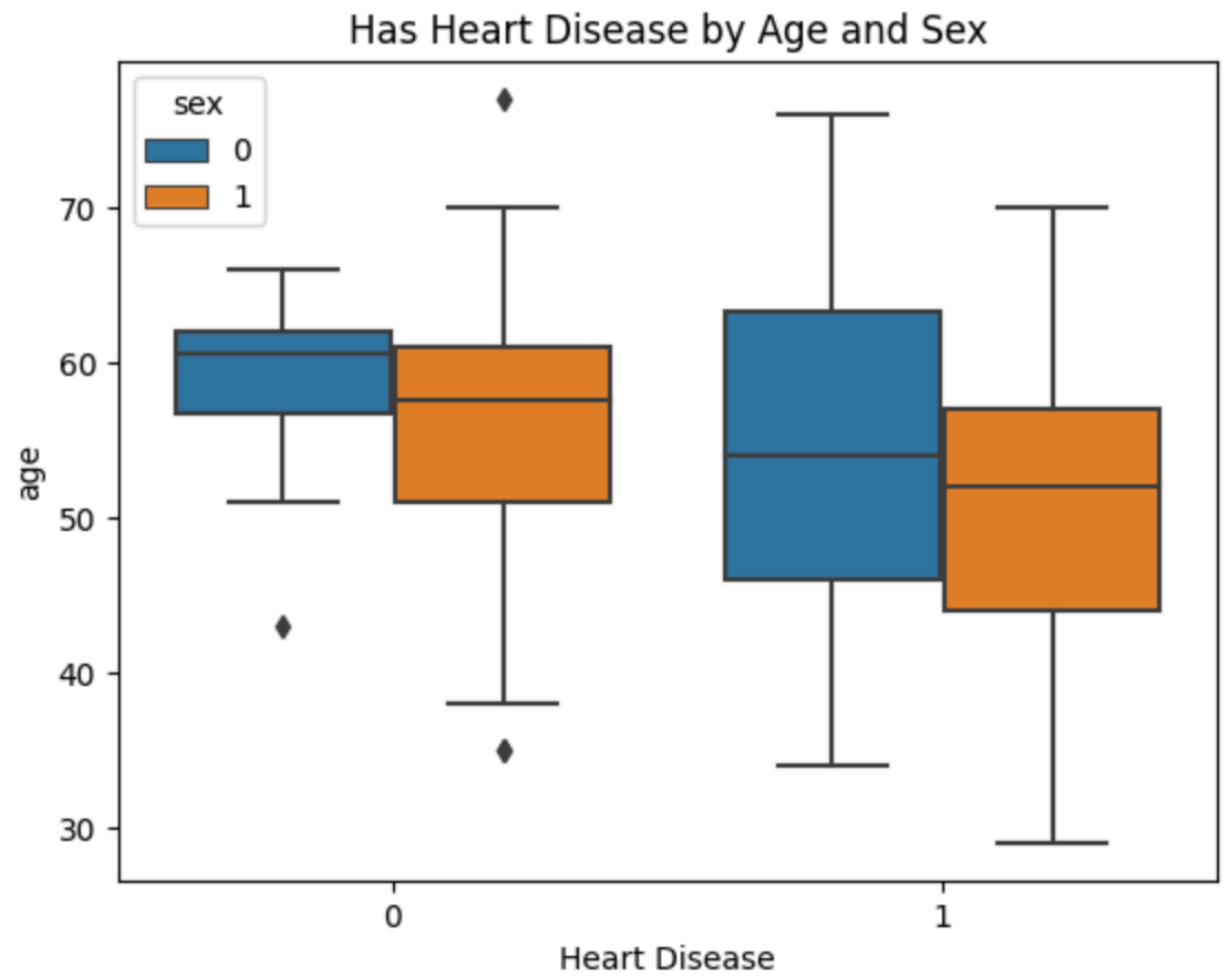
Pfizer Interview

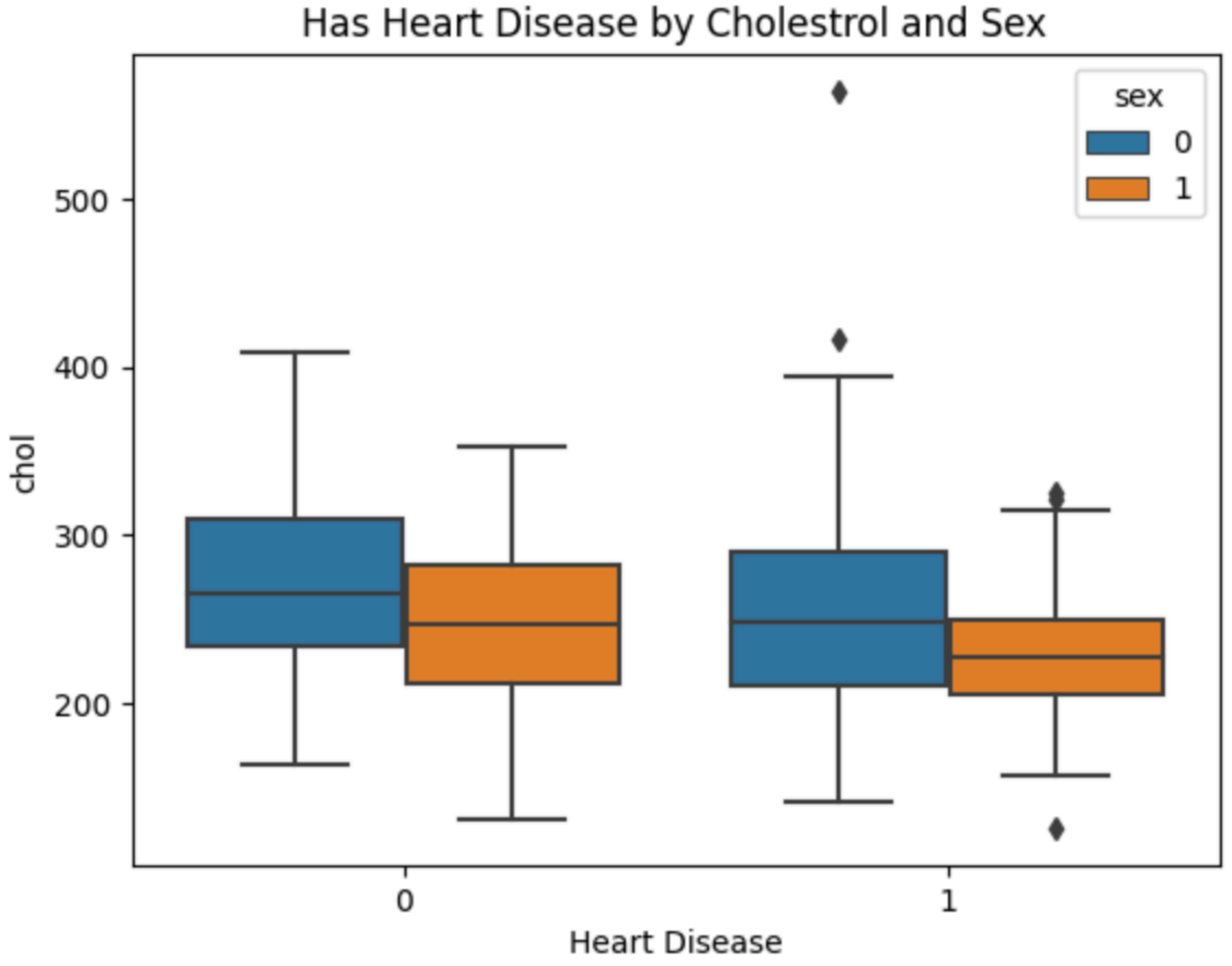
Approach

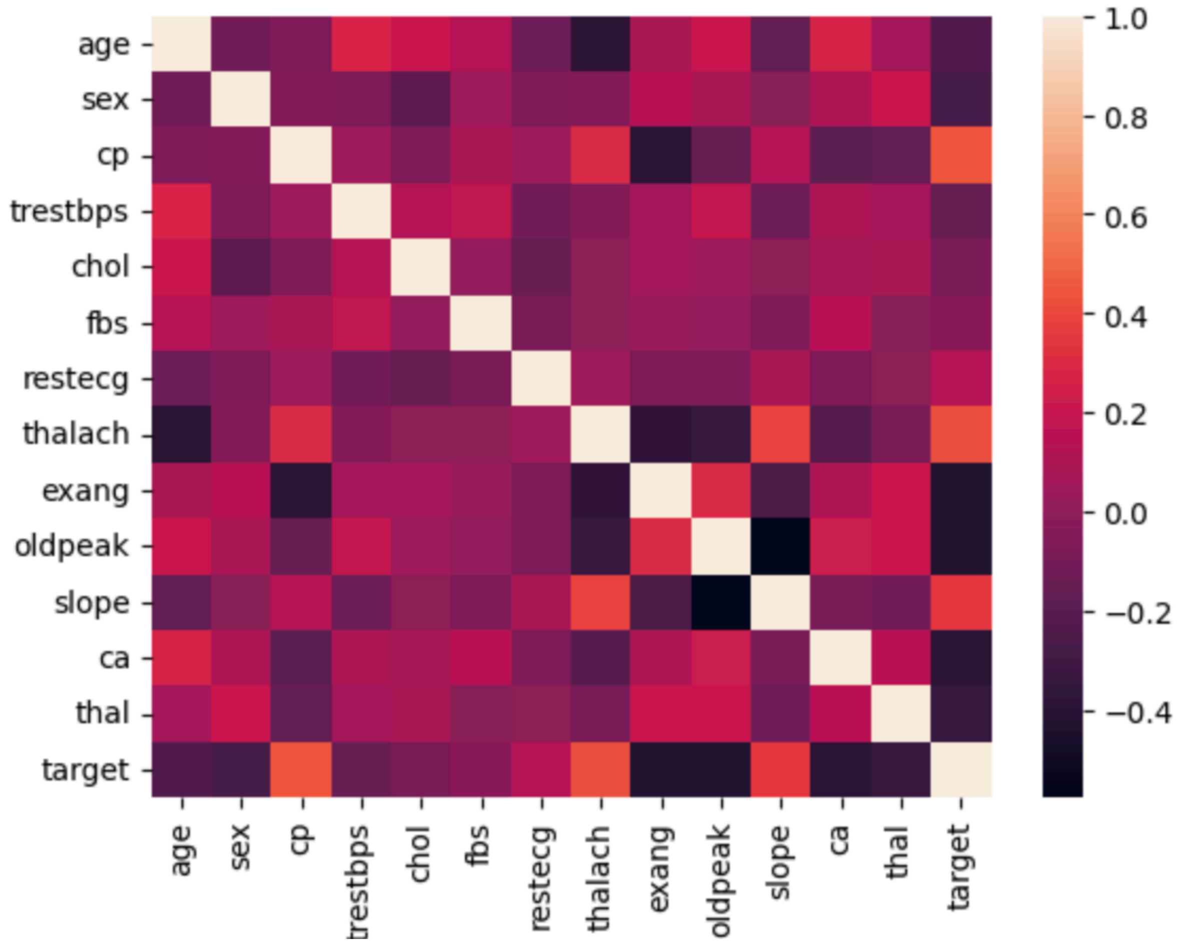
- Conduct exploratory data analysis (EDA)
- Transform Data
  - Create categorical data transformations
  - Standardize numerical data
  - Create interaction effects
- Estimate classification models (Logistic Regression)
    - Use forward selection approach to identify models
    - Short-list models based on model performance
- Summarize findings

Conduct Exploratory Data Analysis (EDA)

- Heart disease patients (in the dataset) are generally younger than healthy patients (across both men and women)
- Heart disease patients have lower cholesterol than healthy patients for men (and for a lesser extent) for women
- Other plots including correlation were made to understand predictors of heart disease

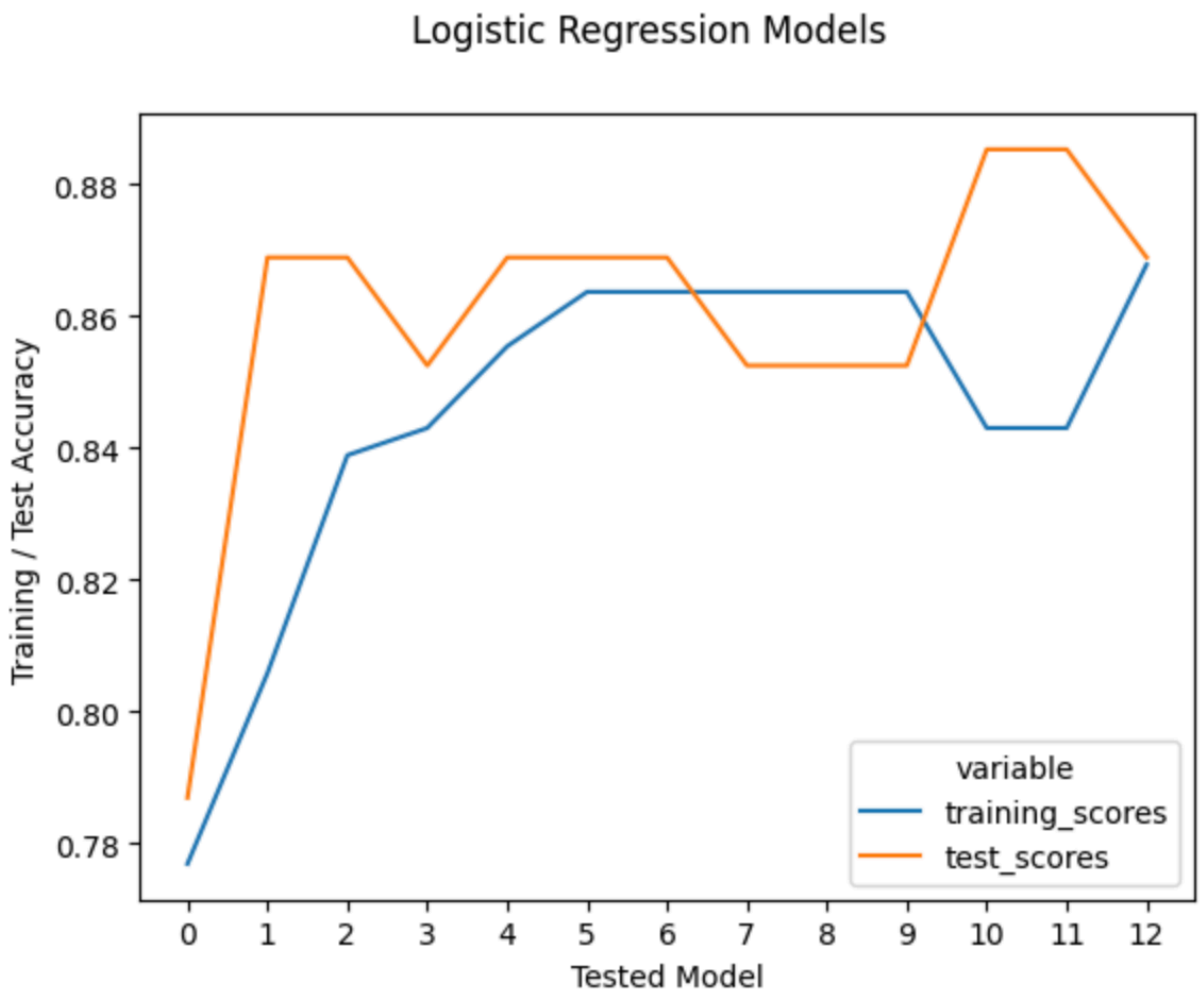

Has Heart Disease by Age and Sex


Has Heart Disease by Cholestrol and Sex



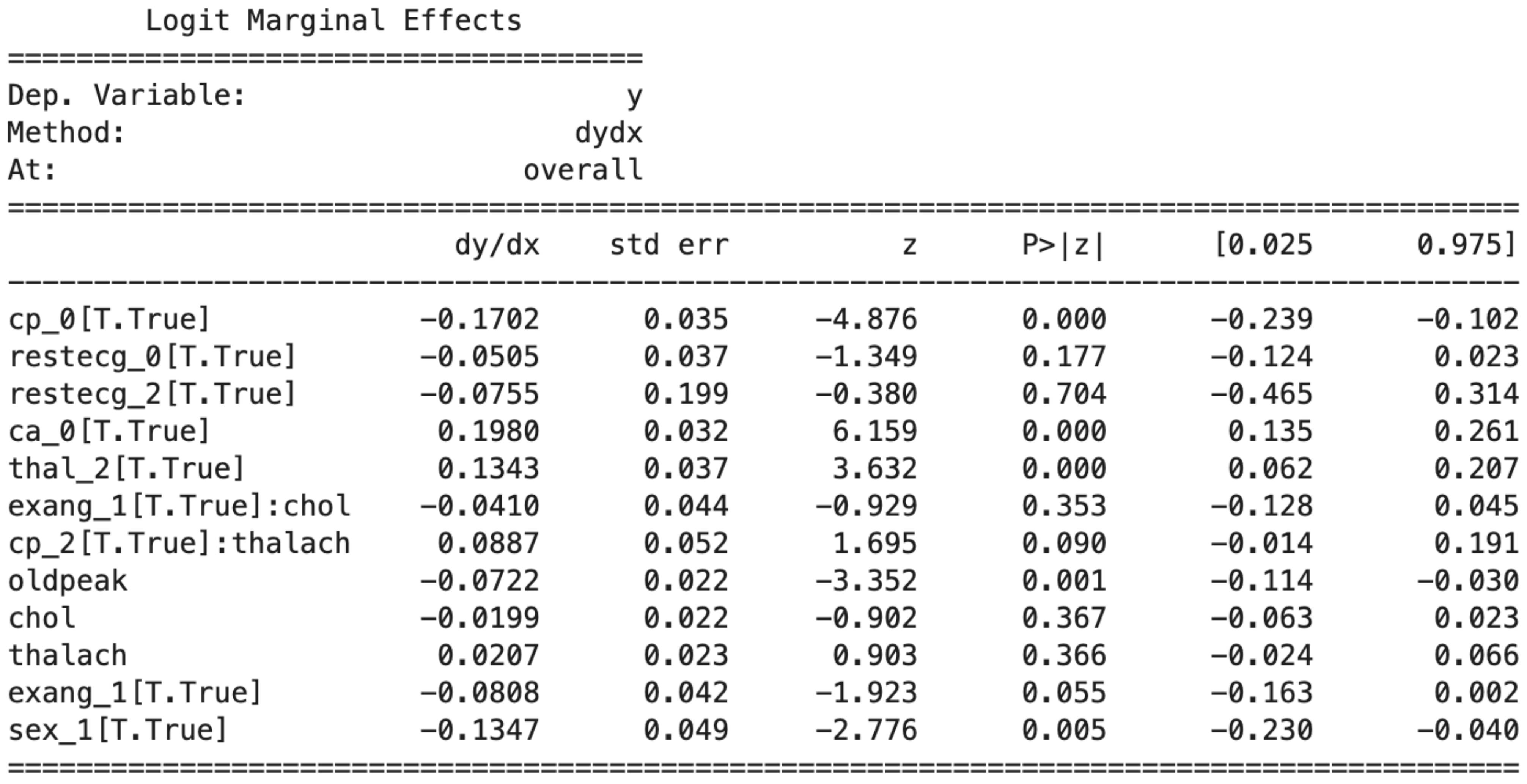
Estimate Classification Models

- Logistic regression was selected due to ease of interpretation
- After data transformations 20% of the dataset was kept as testing dataset and 80% was used for training
- A forward selection process was used to identify which models would perform better.
- Model 11 / 12 have the highest training and test scores (model 12 selected as the final model)

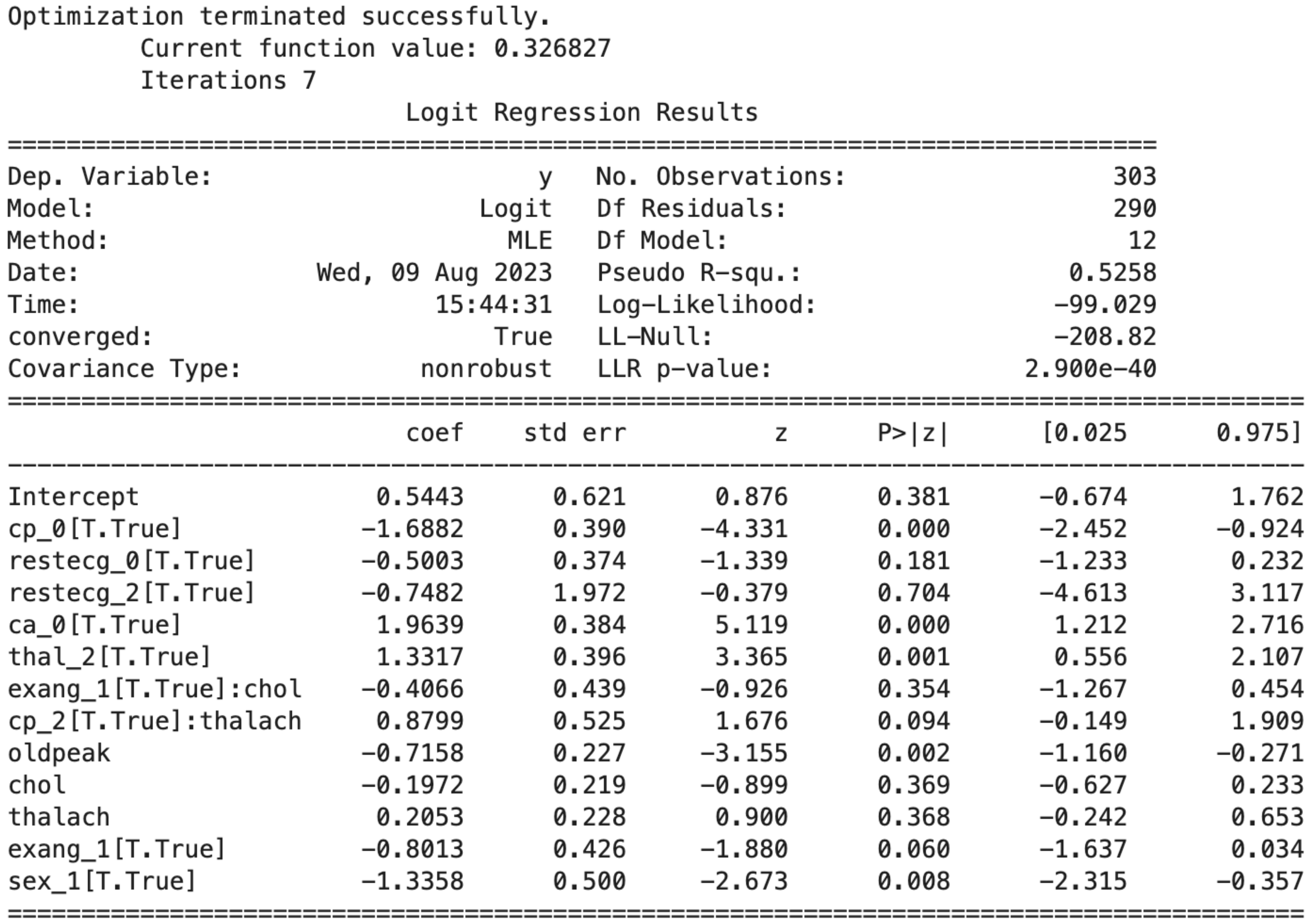


What are the main contributing factors towards heart disease?

- Having zero vessels colored by Fluoroscopy (ca=0), is associated with 19% increase heart disease (all other factors constant)
- Chest pain Type 0 is associated with a 17% reduction in heart disease
- Fixed defects in heart (thal=2) is associated with a 13% increase in heart disease
- Men have a 13% lower risk of heart disease
- 1 standard deviation increase in ST depression is associated with 7% reduction in heart disease
- Exercise induced angina is associated with an 8% decrease in heart disease



```
        Logit Marginal Effects
=====================================
Dep. Variable:                      y
Method:                          dydx
At:                           overall
========================================================================================
                          dy/dx    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
cp_0[T.True]            -0.1702      0.035     -4.876      0.000      -0.239      -0.102
restecg_0[T.True]       -0.0505      0.037     -1.349      0.177      -0.124       0.023
restecg_2[T.True]       -0.0755      0.199     -0.380      0.704      -0.465       0.314
ca_0[T.True]             0.1980      0.032      6.159      0.000       0.135       0.261
thal_2[T.True]           0.1343      0.037      3.632      0.000       0.062       0.207
exang_1[T.True]:chol    -0.0410      0.044     -0.929      0.353      -0.128       0.045
cp_2[T.True]:thalach     0.0887      0.052      1.695      0.090      -0.014       0.191
oldpeak                 -0.0722      0.022     -3.352      0.001      -0.114      -0.030
chol                    -0.0199      0.022     -0.902      0.367      -0.063       0.023
thalach                  0.0207      0.023      0.903      0.366      -0.024       0.066
exang_1[T.True]         -0.0808      0.042     -1.923      0.055      -0.163       0.002
sex_1[T.True]           -0.1347      0.049     -2.776      0.005      -0.230      -0.040
========================================================================================
```



```
Optimization terminated successfully.
         Current function value: 0.326827
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  303
Model:                          Logit   Df Residuals:                      290
Method:                           MLE   Df Model:                           12
Date:                Wed, 09 Aug 2023   Pseudo R-squ.:                  0.5258
Time:                        15:44:31   Log-Likelihood:                -99.029
converged:                       True   LL-Null:                       -208.82
Covariance Type:            nonrobust   LLR p-value:                 2.900e-40
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept                0.5443      0.621      0.876      0.381      -0.674       1.762
cp_0[T.True]            -1.6882      0.390     -4.331      0.000      -2.452      -0.924
restecg_0[T.True]       -0.5003      0.374     -1.339      0.181      -1.233       0.232
restecg_2[T.True]       -0.7482      1.972     -0.379      0.704      -4.613       3.117
ca_0[T.True]             1.9639      0.384      5.119      0.000       1.212       2.716
thal_2[T.True]           1.3317      0.396      3.365      0.001       0.556       2.107
exang_1[T.True]:chol    -0.4066      0.439     -0.926      0.354      -1.267       0.454
cp_2[T.True]:thalach     0.8799      0.525      1.676      0.094      -0.149       1.909
oldpeak                 -0.7158      0.227     -3.155      0.002      -1.160      -0.271
chol                    -0.1972      0.219     -0.899      0.369      -0.627       0.233
thalach                  0.2053      0.228      0.900      0.368      -0.242       0.653
exang_1[T.True]         -0.8013      0.426     -1.880      0.060      -1.637       0.034
sex_1[T.True]           -1.3358      0.500     -2.673      0.008      -2.315      -0.357
========================================================================================
```

Comments: Do the findings make sense?

- The findings are counter-intuitive to expectations and may indicate pecularities of this dataset (and may not be generalizable):
  - In general, women outlive men and have lower risk of heart disease
  - In general, exercise induced pain (angina) is associated with a higher risk of heart disease
  - In general, higher cholesterol is known to increase the risk of heart disease
  - In general, older patients have higher risk of heart disease than younger patients
- Model estimates and analysis suggest the ***opposite*** holds for this particular dataset.
- Therefore the results and findings may not be generalizable across different categories of patients