Week11HW8 Final Report

The dataset included patient medical record information from Cleveland, Hungary, and Switzerland. The dataset is focused on cardiac disease and includes relevant variables like age, history of chest pain, blood pressure, and cardiac testing that included electrocardiograms, exercise testing, and the presence of ST changes. The patients were categorized to have (or not have heart disease). The test file does not include the presence or absence of heart disease and will be used to evaluate the model.

The training dataset has 576 rows with 15 columns. The rows are the patients and the columns are the variables, most of which I previously mentioned.

The dataset is missing a decent amount of information which is likely related to the fact that it's a combination of 3 datasets. For example, cholesterol was listed as "0", which is impossible and likely means the data was not available. In the case of cholesterol, the zeros were converted to NaN.

For missing continuous variables (age, trestbps, thalach, oldpeak) they were imputed with the median of the values and for categorical variables (sex, chest pain, resting ecg, angina, etc) they were imputed with the most frequent value. After imputing, the categorical variables were one-hot encoded and one level was dropped so the logistic regression model had a reference category and avoided perfect multicollinearity.

Continuous variables: age, trestbps, chol, thalach, oldpeak. Categorical variables: sex levels: {0, 1} cp (chest pain type) levels: {1, 2, 3, 4} restecg levels: {0, 1, 2} exang levels: {0, 1} slope levels: {1, 2, 3} ca levels: {0, 1, 2, 3} thal levels: {3, 6, 7} Clinic: {cleveland, hungary, switzerland}

Outliers were identified using the IQR rule for continuous variables. I reviewed them and none of the "outliers" were clinically unreasonable so they were included.

Given the outcome was essentially binary (heart disease or not), I chose to use a logistic regression model. Because I was holding off on using the test set, I split the training set into a development set and a validation set. The split was stratified to ensure both sets had the same proportion of heart disease cases as the complete training dataset. I considered a few alternative models that varied in the number of predictors used in the model—for example, using all predictors versus using only 10 or 15 features determined by sequential forward selection. The reasoning for this was simplicity and practicality. In reality, not all of this clinical information is available for every patient, and if the presence of heart disease can be determined with less data, that would be ideal from a cost perspective.

I chose between candidate models by evaluating their performance on the validation set. Performance was measured primarily by AUC. The model with 10 features ended up having the highest validation AUC (0.929), although the difference between each model was not very large (15 features: 0.927; full model: 0.926). After choosing the model with 10 features, I refit that model on the full training set and reported 5-fold stratified cross-validation AUC/accuracy to ensure the model did not just overperform on that single split.

I tried to keep the model as simple as possible and avoid making it overly complex, unstable, and susceptible to noise or overfitting. I did not use penalized regression as I had already simplified the model by limiting the number of features to 10.

Final Model Performance and Interpretation

metric	train	cv_mean	cv_sd
AUC	0.931	0.906	0.039
Accuracy	0.856	0.825	0.060

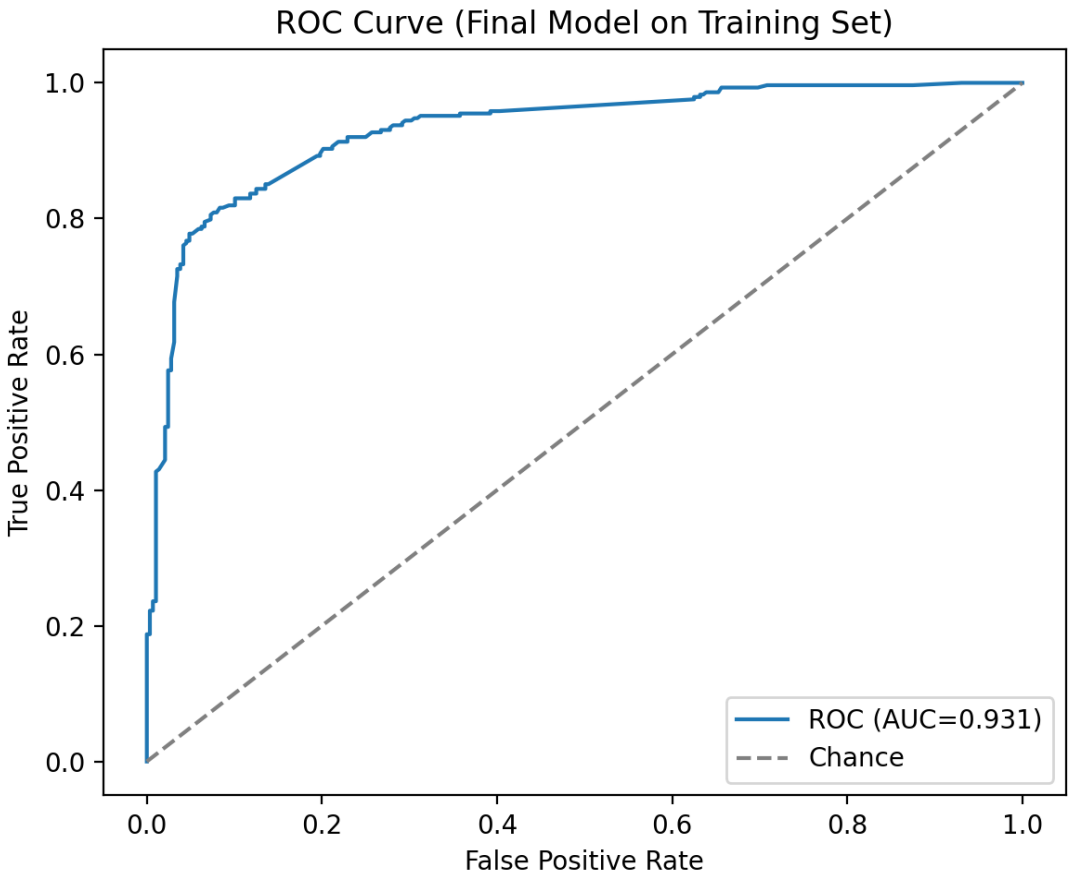
Top effects (odds ratios with 95% CI) from Statsmodels refit:

feature	odds_ratio	ci_low	ci_high
clinic_swit	41.456	15.809	108.713
ca_1	9.042	3.386	24.143
ca_2	7.187	1.745	29.594
cp_4	6.703	3.887	11.560
slope_2	4.879	2.541	9.368
thal_7	3.590	1.857	6.939
sex_1	2.831	1.493	5.370
exang_1	2.816	1.546	5.129
fbs_1	2.404	0.966	5.982
oldpeak	2.039	1.508	2.758

Overall, the strongest predictors of heart disease in the final model were related to cardiac testing/exercise findings and clinical severity indicators. For example, ST-segment depression had an odds ratio of about 2.0, meaning higher values were associated with higher odds of heart disease. This makes sense because it's a common indicator for ischemic changes to the myocardium. In addition, exercise-induced angina and being male were also associated with increased odds (OR ~2.8 for both). Chest pain type 4 showed a large association with heart disease (OR ~6.7), which makes clinical sense because more severe/typical chest pain patterns are more likely to indicate disease.

Interestingly, the Switzerland clinic indicator had a very large odds ratio (~41). I don't think this means Switzerland itself causes heart disease, but it might reflect that the Switzerland records were collected differently (different patient mix and a lot more missing data), which can make the clinic variable act like a "site marker" rather than a purely biological predictor.

Figure 1. ROC curve for the final model (training set). The x-axis is the false positive rate and the y-axis is the true positive rate.


ROC Curve (Final Model on Training Set)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC (AUC=0.931)
Chance