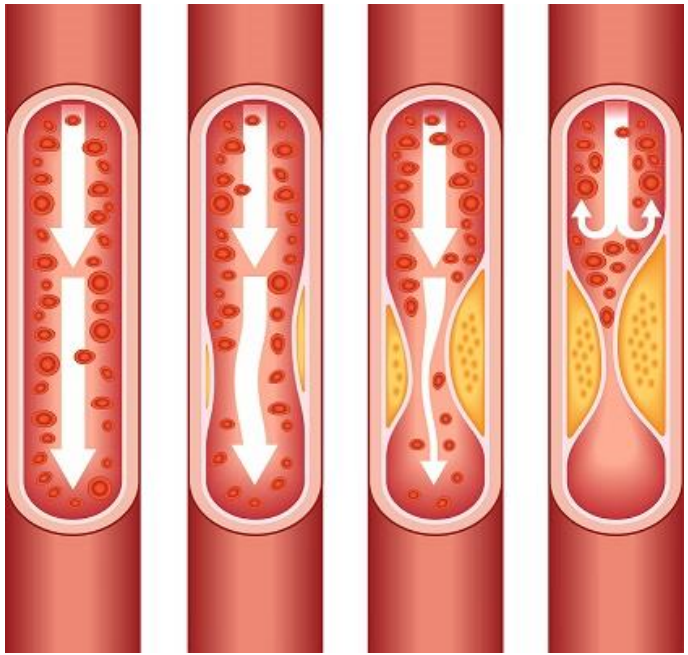
Cardiovascular Disease Prediction

BE 188: Machine Learning and Data-Driven Modeling in Bioengineering

Nilay Shah, Aditya Gorla

Introduction

- Cardiovascular disease(CVD) is number one cause of death worldwide and leads to more than 600,000 deaths in the United States annually
- CVD healthcare is still very reactive
- We would like it to be proactive



Data

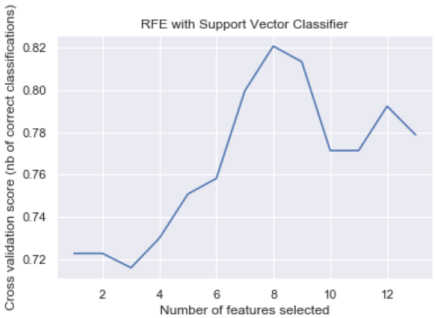
- CVD dataset was obtained from a challenge hosted by DrivenData
    - It was compiled from the Cleveland Heart Disease Database
- Consists of a subset of 14 attributes from 180 patients
    - Has ordinal, numerical and categorical data

Key Model Evaluation Metrics

1. **Precision:** ∑ True Positive/ ∑ Predicted Positive (i.e. the ability of the classifier to not label a sample as positive if it is negative)
2. **Recall:** True Positive rate or Sensitivity. ∑ True Positive/ ∑ All Positive (i.e the ability of the classifier to find all the positive samples)
3. **Log Loss:** measures the performance of a classification model's predictive ability by penalizing the predicted probability as it diverges from the actual label
4. **ROC AUC:** quantifies the diagnostic ability of a binary classifier system as its discrimination threshold is varied

Feature Selection

- Recursive feature elimination with cross-validation (RFECV)
- Run with both logistic regression model and linear support vector classifier
- Produced 6 common features and 3 variable features



Parameter Optimization

- Created a powerset of the 3 variable features and combined each with the 6 common ones
- Determined optimal combination of features for each classifier
- Used GridSearchCV to exhaustively search parameter values with cross-validation

Model Building

- Built logistic regression and SVM models with the best no. of features for each and their optimized parameters
- Logistic Regression:

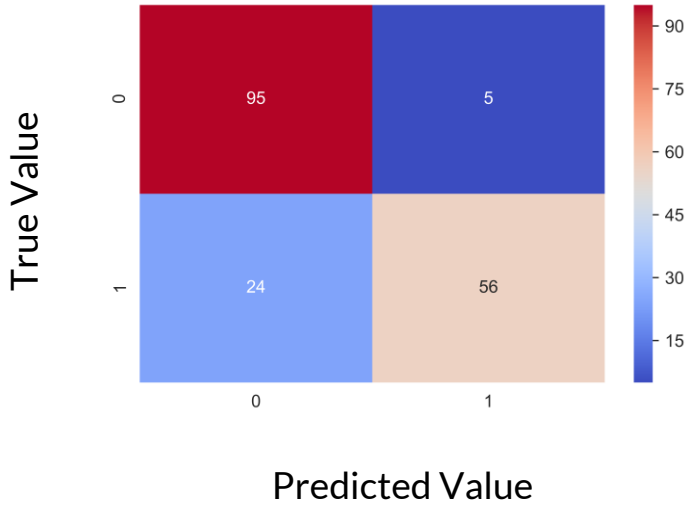
  {'logistic__C': 0.02, 'logistic__penalty': 'l2', 'logistic__solver': 'newton-cg'}
- SVM Model:

  {'svc__C': 0.05, 'svc__degree': 3, 'svc__gamma': 'auto', 'svc__kernel': 'poly'}

K-fold Cross-Validation

- Ran stratified 80-fold Cross-Validation
- Stratification ensures each fold contained roughly the same proportions of the two types of class labels

Logistic Regressions Results

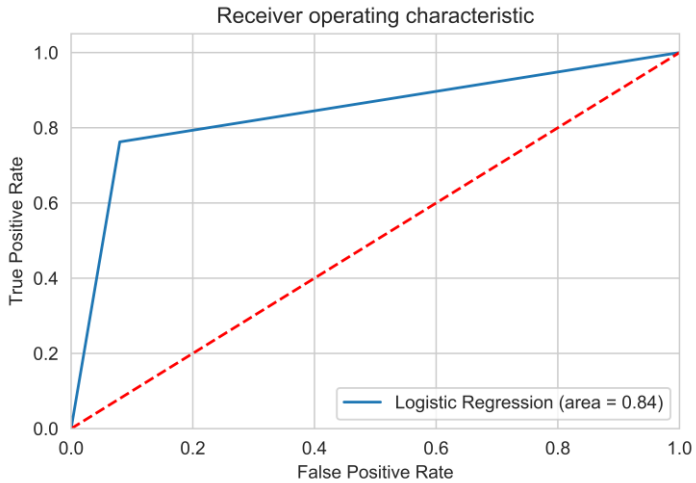


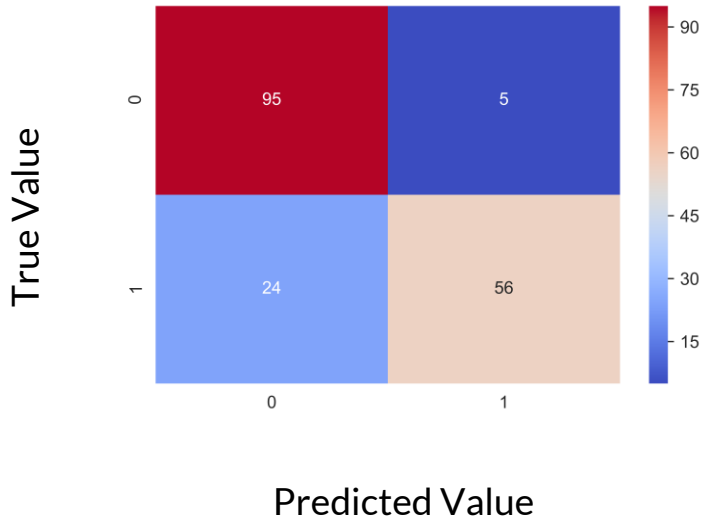
Data obtained from 80-fold Stratified Cross Validation

Log loss: 0.458

	precision	recall	f1-score	support
0	0.80	0.95	0.87	100
1	0.92	0.70	0.79	80
micro avg	0.84	0.84	0.84	180
macro avg	0.86	0.82	0.83	180
weighted avg	0.85	0.84	0.84	180

SVC Results





Data obtained from 80-fold Stratified Cross Validation

	precision	recall	f1-score	support
0	0.83	0.92	0.87	100
1	0.88	0.76	0.82	80
micro avg	0.85	0.85	0.85	180
macro avg	0.86	0.84	0.85	180
weighted avg	0.85	0.85	0.85	180

Conclusion/Future Directions

- Log Loss of 0.42277 achieved with best model submitted to competition - top 21% of submissions
    - Continue to optimize model to achieve log loss ~0.25
- Logistic Regression model minimizes false positives, while SVM model minimizes false negatives
- Considering combining dataset with predictions from logistic regression and SVM models to train boosted decision trees to improve Sensitivity and Specificity