Predicting Cardiovascular Disease

Aditya Gorla & Nilay Shah

Aditya Gorla:

Project: Data organization, feature selection, model testing, and cross-validation

Report: Introduction, problem definition, results

Nilay Shah:

Project: Data exploration, feature selection, parameter optimization, model testing, and submission to contest

Report: Introduction, methods, general editing

Introduction

Cardiovascular disease (CVD) is the number one cause of death worldwide and leads to more than 600,000 deaths in the United States annually. However, the current state of CVD care is often reactive since patients do not seek treatment until the disease has progressed to the point where it necessitates various medical procedures or even surgery. Developing a computational model for predicting the presence of heart disease based on physiological variables gathered from patients could prove to be a valuable tool in the research and prevention of this disease. Earlier detection could help patients manage their conditions with only medications and lifestyle changes. Tools such as these will help transition healthcare to being more proactive, where the goal is mitigation before the heart attack or stroke occurs rather than treatment after the damage has already happened. In this exploratory project, we will work to develop a model for predicting the probability that a given patient has heart disease. Our dataset will obtained from the challenge hosted by DrivenData and will consist of a subset of 13 attributes from 180 patients, assembled from the Cleveland Heart Disease Database. This dataset has been used in previous research papers studying heart disease diagnosis. Feshki et al[1] have used this dataset to demonstrate predictive accuracy of 91.94% with their model. These papers can serve as benchmark for us as we seek to develop a more optimized algorithm.

Problem Definition

The primary goal of this project is to develop a classifier (or a group of classifiers) which can predict whether a person has a heart disease or not. To evaluate the performance of our model we shall use 4 key metrics: 1) Precision: ∑ True Positive/ ∑ Predicted Positive, 2) Recall: ∑ True Positive/ ∑ All Positive (i.e. Sensitivity), 3) Log Loss: a measure of a classification model's predictive ability that penalizes the predicted probability as it diverges from the actual label and 4) Area Under the Receiver Operating Characteristics curve (AUROC): a value that quantifies the discrimination ability of a binary classifier. These metrics will be evaluated under stratified K-fold cross validation.

The Cleveland Heart Disease Database dataset is relatively manageable, with only 13 key parameters for analysis. The process of developing a classifier which can accurately predict heart disease will give us an insight into which physiological features are the most correlated to the development of heart disease. This will allow researchers to focus their efforts into better understanding the relation between the selected features and the causes of heart disease. Finally, a highly accurate classifier could one day be implemented in the field to idenify patients that are at high risk for heart disease before they've suffered a stroke or heart attack, opening the door to a fundamentally proactive way of practicing medicine.

Methods

Data Organization and Exploration

The training data was loaded and organized into a pandas dataframe and separated into the input variables, 13 various physiological features, and the output variable, a binary value for whether or not the patient has heart disease. The 13 features consist of binary, numerical, ordinal, and categorical data. The thallium stress test feature measuring blood flow to the heart, was given as a categorical variable and in order for it to be used in our classifiers we encoded the given values of normal, fixed defect, and reversible defect as 0, 1, or 2 respectively. This data was provided for a total of 180 patients, 100 of which displayed no heart disease and 80 which did, indicating that we have fairly balanced classes with a ratio of 5:4.

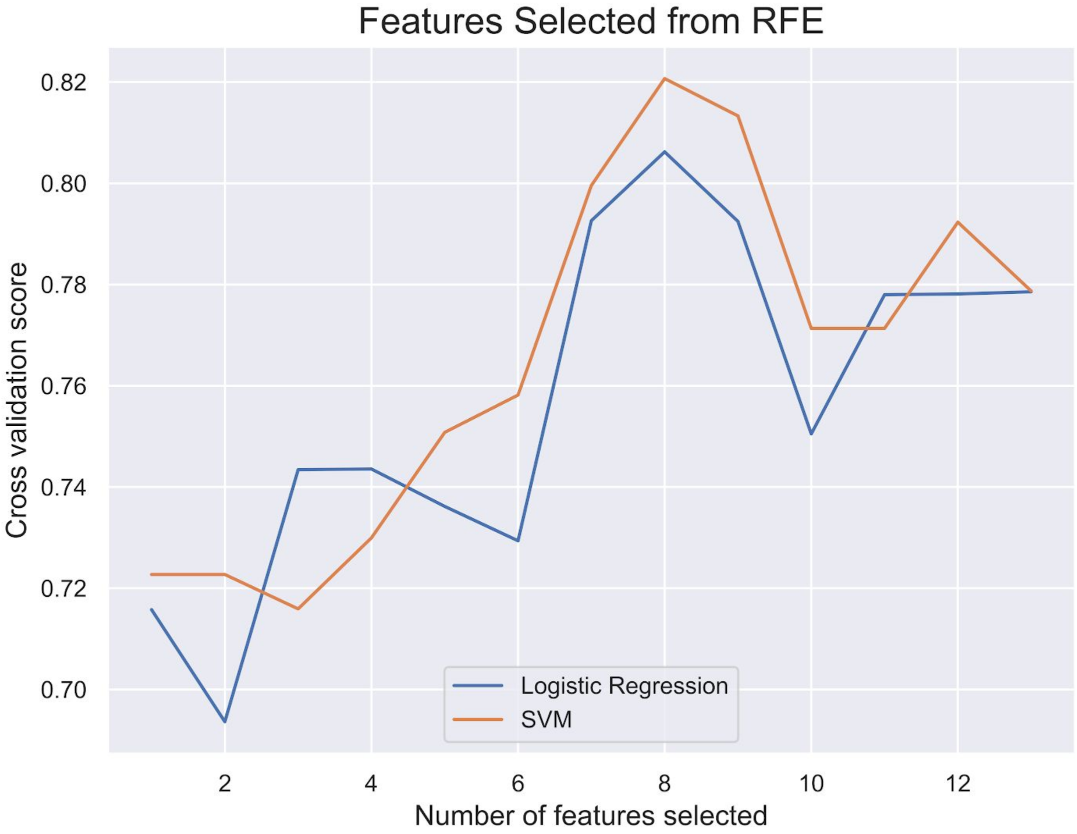
For the initial exploration of the data, we examined the average of each of the features as organized by the output variable, the presence of heart disease. For some variables, such as chest pain induced by exercise, there was a clear separation between the two classes with an average of 0.13 for those without heart disease and 0.55 for those with the disease. On the other hand, there was not as prominent a difference in other features such as resting blood pressure, where the values were 130.12 for those without heart disease and 130.80 for those with the disease.

Feature Selection

We obtained baseline performance values with the entire dataset by building logistic regression and support vector machine (SVM) classifiers through the algorithms available in Python's scikit-learn library. However, based on our observations from the mean values, we believed our models could be optimized through feature selection.

Regularization through ridge regression or LASSO regression is built into the logistic regression function as the penalty parameter, so to further trim our dataset we implemented recursive feature elimination with cross validation (RFECV). This method works by taking in a classifier and optimizing it by doing further variable selection. The given classifier is trained on the initial set of features and the beta coefficients are examined to determine the importance of each feature. The closer the value is to 0, the less correlation there is between that feature and the output variable. The least important of these features are recursively pruned and the model is trained on smaller and smaller subsets of the features. 5-fold cross validation is performed in conjunction with this process to determine the optimal number of features to select which ensures that we do not overfit the selected features to our training data.

We ran RFECV with both logistic regression and SVM classifiers (SVC), and in both cases it was determined that eight features did the best in minimizing prediction error by giving the highest cross validation score. In addition, both classifiers agreed on the same eight features as being the most relevant. This gives us added certainty that one classifier isn't overfitting to the data. The figure below is a cross validation scores vs. feature index plot, for each our classifiers.



The eight key features as selected by RFECV and the description of each is as follows:

1. **Slope_of_peak_exercise_st_segment:** the slope of the peak exercise ST segment, an electrocardiography read out indicating quality of blood flow to the heart
2. **Thal**: results of thallium stress test measuring blood flow to the heart
3. **Chest_pain_type:** ordinal measurement of chest pain type (4 values)
4. **Num_major_vessels:** number of major vessels (0-3) colored by flourosopy

5. **Fasting_blood_sugar_gt_120_mg_per_dl:** binary measurement of whether their fasting blood sugar > 120 mg/dl
6. **Oldpeak_eq_st_depression:** oldpeak = ST depression induced by exercise relative to rest, a measure of abnormality in electrocardiograms
7. **Sex**
8. **Exercise_induced_angina:** binary measurement of exercise-induced chest pain

Parameter Optimization

To further improve the performance of each of our two models, we selected optimized parameters for each classifier by performing an exhaustive cross-validated grid-search over select specified values (GridSearchCV).

For the logistic regression classifier, we optimized for the C value (an inverse of regularization strength), testing values from 0.0001 to 10, the penalty, or the type of regularization performed (l1 - Lasso Regression, l2 - Ridge Regression) and the solver used (newton-cg, lbfgs, sag, saga). The optimal parameters were determined to be the newton-cg solver with a l2 penalty and a C-value of 0.05.

For the SVM classifier, we also optimized for the C value, testing values from 0.0001 to 10, the kernel used (poly, rbf, sigmoid), the degree (allowed to range from 2-9), and the gamma value (auto or scale). The optimal parameters were determined to be a polynomial kernel with a degree of 2, a C-value of 0.5 and gamma set to auto.

Building the Models and K-fold Cross Validation

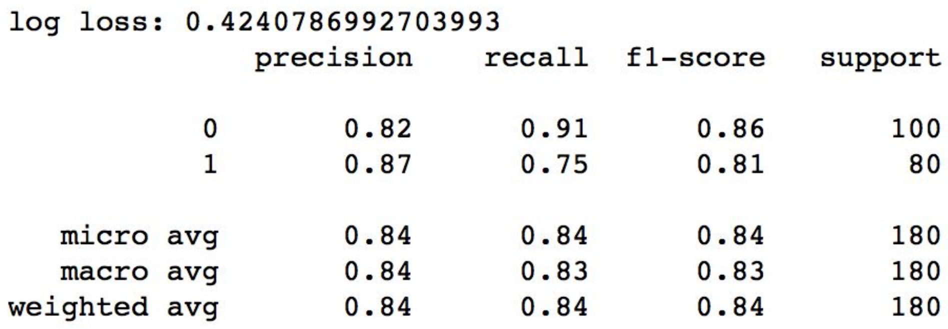
A logistic regression model and SVM model were built with the parameters determined from the GridSearch and the features selected by RFECV. 80-fold Stratified cross validation was performed on each of the models in order to evaluate their predictive capabilities. Stratified cross-validation was employed to ensure that each fold contained roughly the same proportions of the two types of class labels.

Results and Future Directions

Logistic Regression Results

Using the optimized logistic regression parameters we get the following performance metrics results.

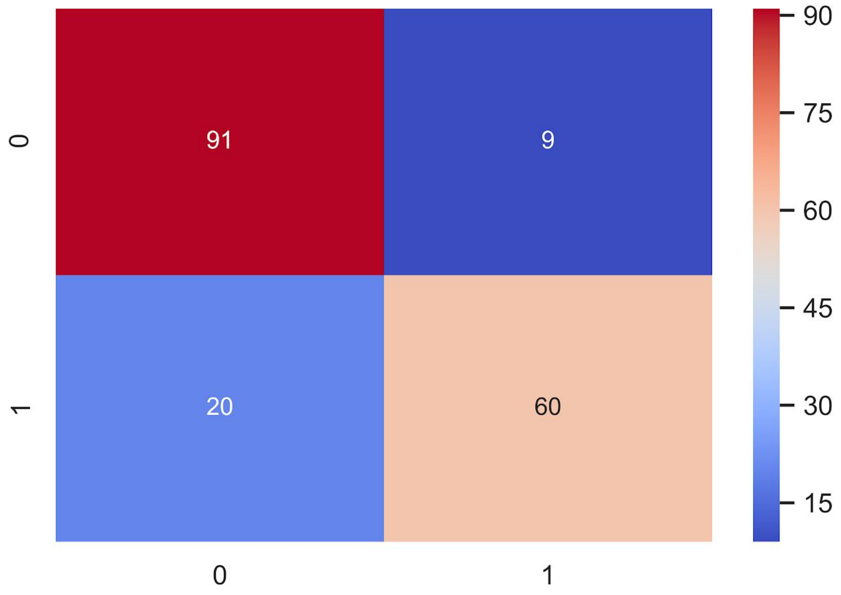
NOTE: The values in the vertical axis of the heatplot refer to the true states and the values in the horizontal axis refer to the predicted values.

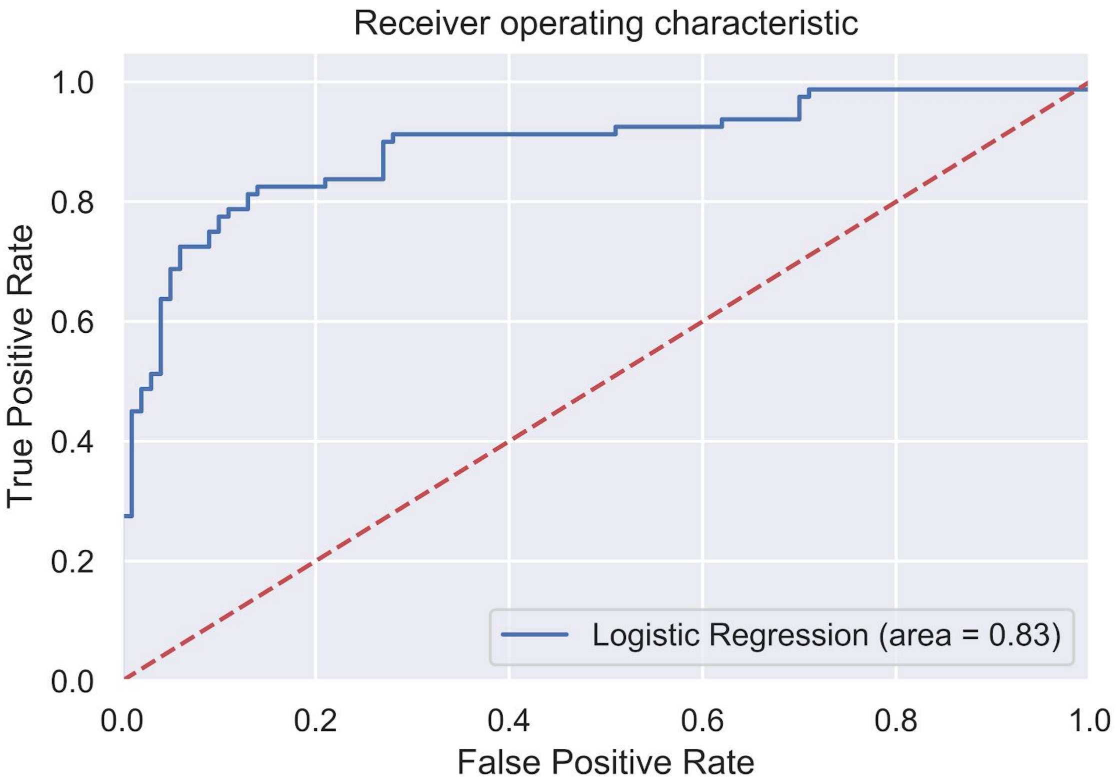


```
log loss: 0.4240786992703993
              precision    recall  f1-score   support

           0       0.82      0.91      0.86       100
           1       0.87      0.75      0.81        80

   micro avg       0.84      0.84      0.84       180
   macro avg       0.84      0.83      0.83       180
weighted avg       0.84      0.84      0.84       180
```

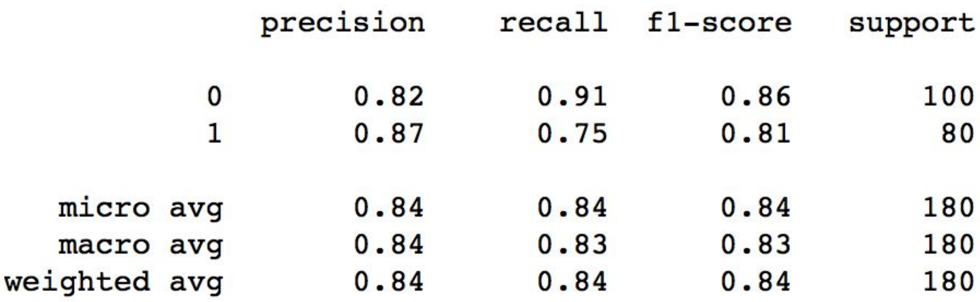




SVM Classifier Results

Using the optimized support vector classifier parameters we get the following performance metrics results.

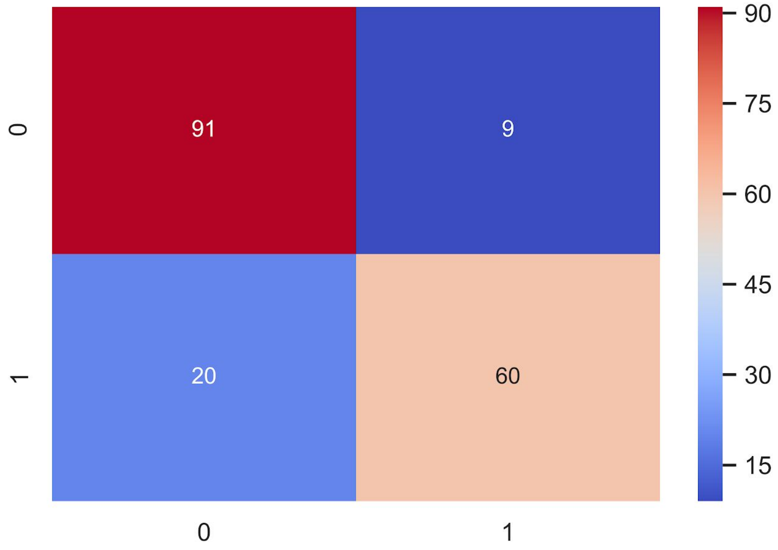
NOTE: The values in the vertical axis of the heatplot refer to the true states and the values in the horizontal axis refer to the predicted values.

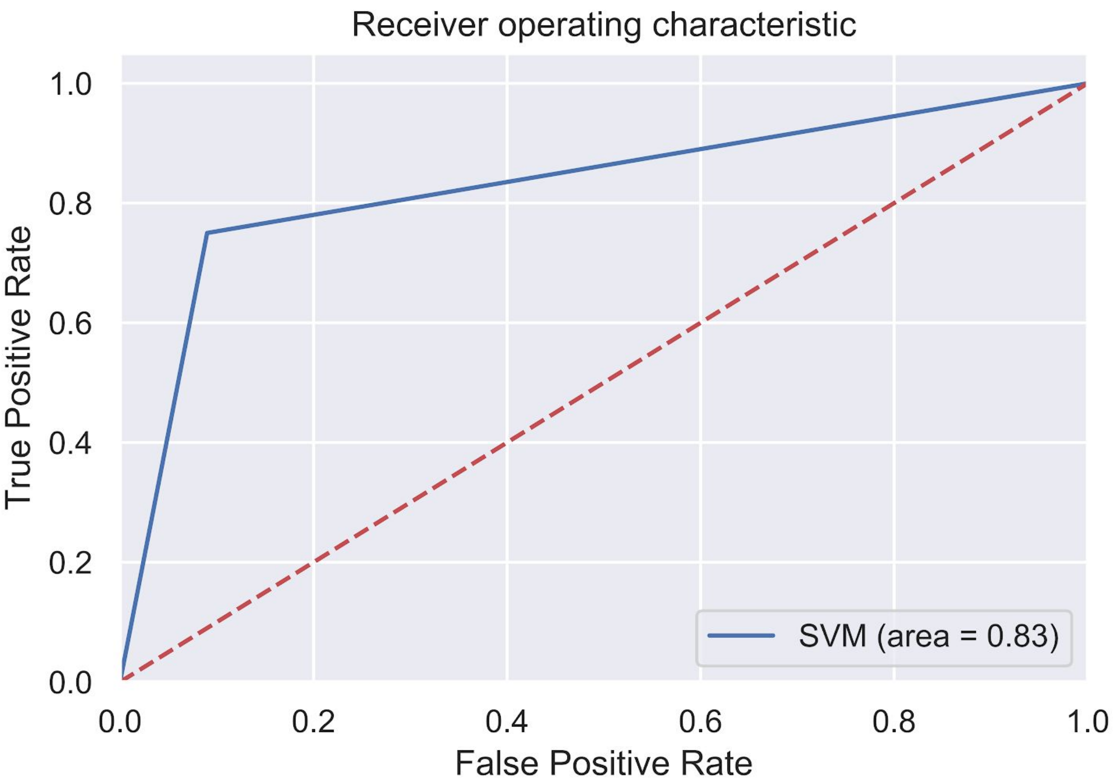


```
              precision    recall  f1-score   support

           0       0.82      0.91      0.86       100
           1       0.87      0.75      0.81        80

   micro avg       0.84      0.84      0.84       180
   macro avg       0.84      0.83      0.83       180
weighted avg       0.84      0.84      0.84       180
```





Discussion

From the above results, we see that the precision, recall and confusion matrix values are exactly the same for both the logistic regression and SVM classifiers. Next, we notice that the AUROC for both the methods are also the same. These similarities give us confidence in the fidelity of our model since the results from different classifiers seem to converge. However, it also indicates that we may be stuck at a local minima with logistic regression and SVC. The predictions generated from each model when compared to the true heart disease status seem to misclassify the same data. This informs us that these data points have values that are sufficiently more convoluted than those which can be resolved by logistic regression or support vector machines. In the future, we will need to explore other classification techniques for these confounding data points to further improve our accuracy.

The next key takeaway from this model is the discovery of the eight key features from the dataset. The two models independently concluded that the same eight features had the most relevance for predicting CVD. The beta coefficients from the optimized logistic regression classifier were the highest for num_major_vessels, thal, and chest_pain_type with values of 0.313, 0.280, and 0.279 respectively, indicating that of the eight features, these three demonstrated the greatest correlation for predicting heart disease in our model. This seems valid considering that all three are major indicators for atherosclerosis, the progression of which leads to the development of CVD. Beta coefficients are not available for polynomial kernel SVM classifiers, but we can still say with a high degree of confidence that the eight selected features, with those three in particular, require further clinical research. Researchers should spend time further understanding the physiological links between these features and cardiovascular disease.

Finally, our logistical regression model gave us a log loss of 0.424 during cross validation. However, when we used the same optimized logistic regression model to generate results for the competition test data, for which we don't have the true labels, we achieved a log loss of 0.404. This puts us in the to 80th percentile of the teams competing the in competition.

Future Direction

There is still much work to be done in making an effective classifier for CVD. We will most certainly continue working on this project to optimize our model in the future since there are still seven months left in the competition. Our next steps for this project include:

1. Exploring other classification techniques to boost our precision and recall
2. Consider combining the eight key feature data set and predictions from the logistic regression and SVC into a new data set and running the new data set through a boosted decision tree
3. If boosted decision trees seem promising, building a random forest of boosted decision trees and evaluating performance
4. Use these methods to achieve a log loss <= 0.25 which would put us in the top 10 in the competition

References

1. Feshki, M. G. & Shijani, O. S. Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network. in 2016 Artificial Intelligence and Robotics (IRANOPEN) (IEEE, 2016). doi:10.1109/rios.2016.7529489