Name: Vaishnavi Gannavaram

Instructor: Dr. Srikanth Mudigonda

Course: 5330-01: Predictive Modelling and Machine Learning

FINAL PROJECT REPORT

CLASSIFICATION DATA SET

DATA SET 1) STATLOG (HEART) DATA SET - Classification

1 Introduction

1.1 Motivation

1. **Why have you chosen to analyze the dataset you have chosen?**

In this industrialized society, “heart disease” is one of the main causes of most fatalities. Heart disease can be brought on by living a sedentary lifestyle, smoking, drinking, and consuming an excessive amount of fat, etc., which can increase blood pressure. Factors causing the disease can be many but, the timely diagnosis of an illness is crucial to its care. Therefore, Using Machine Learning approaches that can aid in the early detection and prediction of cardiac disease would be extremely helpful for the Health Care Sector.

The motivation for the study is to find the most efficient ML algorithm for the detection of heart diseases.

Also, finding the aggregated relative importance of predictors across all models and drawing appropriate practical conclusions from that which can help health care providers is the second goal.

1.2 Source of the data

2. **What is the source of the data?**

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Also, found in Kaggle: Statlog (Heart) Data Set | Kaggle

3. **If the dataset is public, provide its URL; if the dataset is proprietary (from your research/workplace), provide a brief overview of the context in which the dataset was collected.**

This is an open data set which is clean and ready to use. Present in Kaggle and in UCI Machine Learning Repository. This dataset is publicly available for research purposes only.

Cite at:
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

1.3 Details of the dataset

Provide details, via a table with appropriate columns, of the variables from the dataset you considered predictors and outcome(s). For each variable, provide its name, its measurement type, a brief description, and its role (predictor and/or outcome).

S.No.	Variable	Description	Distinct Value	Type
1	Age	Patient age in year	29 to 77	Numerical
2	Sex	Gender	0= female, 1=male	Binary
3	CP	Chest pain type	1= typical angina, 2= atypical angina, 3= non anginal pain 4= asymptomatic	Nominal

4	trestbps	Resting blood pressure in mmHg	94 to 200	Numerical
5	Chol	Cholesterol in mg/dl	126 to 564	Numerical
6	fbs	Fasting blood sugar in patient > 120mg/dL	0= false 1= true	Binary
7	restecg	Resting ECG results	0=normal 1=having ST-T wave abnormality 2= LV Hypertrophy	Nominal
8	thalach	Maximum Heart rate achieved	71 to 200	Numerical
9	exang	Used to identify if there is an exercise induced angina. If yes=1 or else no=0	0= no 1= yes	Binary
10	oldpeak	describes patients' depression level.	0 to 6.2	Numerical
11	slope	describes patient condition during peak exercise. It is divided into three segments (Unsloping, Flat, Down sloping)	1= upsloping 2= flat 3= down sloping	Nominal
12	ca	Result of fluoroscopy.	0,1,2,3	Nominal
13	thal	Defect type	3= normal 6= fixed defect 7= reversible defect	Nominal
14	presence	Heart disease present or absent	0= absence 1= presence	Binary

2 Analyses

2.1 Methods of analysis

4. **What are the appropriate modeling techniques to build predictive models of the outcome?**

To build predictive models of the outcome, ML algorithms such as Random Forest, Linear Discriminant Analysis, Quadrant Discriminant Analysis, Logistic Regression, and Naive Bayes classification techniques are appropriate.

5. **Of the techniques that are appropriate, which ones have you used? What is the rationale for using them?**

- The **Bayes rule** is the foundation of the Naive Bayes algorithm []. The primary presumption and most crucial factor in classifying a dataset is its attribute independence. It is quick and simple to forecast and works best when the independence presumption is true. Equation 1 illustrates how the Bayes theorem determines the posterior probability of an event (A) given some prior probability of an event (B), denoted by P(A/B) [10]
- Regression and classification both employ **Random Forest methods**. The data is organized into a tree, and predictions are based on that tree. Even with a substantial number of record values

missing, the Random Forest algorithm can still produce the same results when applied to huge datasets. The decision tree's generated samples can be preserved and used to different sets of data.

- A classification approach known as **logistic regression** is usually applied for binary classification tasks. Because there are 13 independent factors, logistic regression is effective for categorizing data.
- Being both a classifier and a dimensionality reduction method, linear discriminant analysis **(LDA)** is quite popular.
- Quadratic discriminant analysis, or **QDA**, is an extension of linear discriminant analysis. The observations from each class are assumed to be normally distributed in this approach, which is like LDA, but it does not assume that each class has the same covariance matrix.
- Hence, these techniques are thought to be appropriate for this task.

6. **Which techniques have you excluded in building predictive models? What is the rationale for excluding them?**

As KNN, RIDGE, and LASSO are used under regression analysis those techniques are excluded.

7. **Which metric(s) have you used for evaluating the predictive performance of the models? Why is/are it/they appropriate?**

The metrics prediction and Accuracy score are used to analyze the algorithm's performance.

8. **Are any pre-processing steps required? If so, identify and explain what they are. If not, explain why no pre-processing steps are needed.**

Centering and Scaling were done.

PCA – YES, AIC of the original total model is 207.6 and the first 10 principal component model is 205.52, #Although the pc model performed better, the AIC difference is very slight. Therefore, I feel, we can use either model in the further steps.

CLUSTERING – NO, not necessary

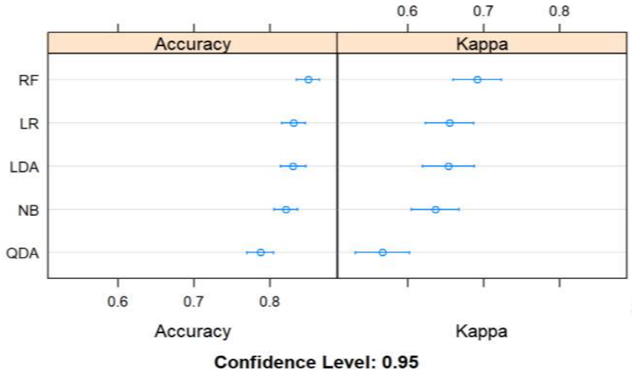
Training set = 75%, Holdout set = 25%

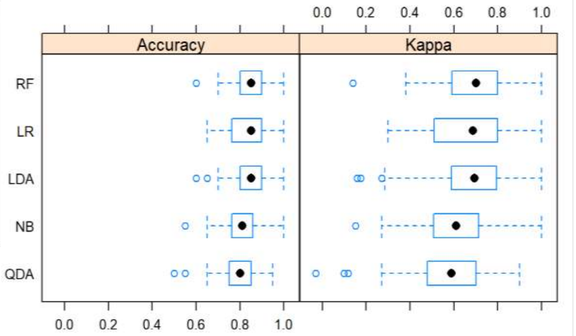
2.2 Results and interpretation

9. **Present a summary of the results of the predictive ability of all the techniques you have used, in the form of a table with appropriate columns that capture the information needed to represent details of a model's predictive performance. Use one row per each modeling technique.**

Model ranking	**Modelling technique**	**Model accuracy**	**Model kappa**	**pred. Accuracy**
1	**rf_model (random forest)**	**85.003**	**0.691**	**92.53**
2	**Logisticfit (logistic regression)**	**83.102**	**0.655**	**89.55**
3	**Ldafit (LDA)**	**83.038**	**0.653**	**88.06**
5	**QdaFit (QDA)**	**78.714**	**0.566**	**83.58**
4	**NBfit (naïve bayes)**	**82.369**	**0.6422**	**89.55**

10. **Complement the numerical summary above with appropriate graphs that depict the variability in the different model's performance. Use the graphs from the caret example https://topepo.github.io/caret/model-training-and-tuning.html#an-example - the graphs are created using bwplot() and dotplot() functions for depicting comparisons of different models' performance measures.**





11. **How would you compare the models based on their predictive performance, along with other information that needs to be considered?**

The Random Forest model ranks first among all the models in terms of accuracy and kappa values with 92.53 predictive accuracy d 0.691 kappa value.

While the logistic regression fit and Linear discriminant analysis fit are closely the same. And the QDA model ranks last among all the other models.

12. **Provide a ranking of the models based on their predictive performance. This can be provided as part of the table mentioned in point 1. above.**

Mentioned in the table above.

13. **Combining the results from all the models provide a ranking, using appropriate evidence, of the predictor variables you have used in the models.**

Noted the variable importance of all the models using varImp () and then computed the average importance of each variable and ranked them according to that average importance.

Ranking order of predictors according to aggregate variable importance of all models:

Ca-92.05, **cp**-90.134, **thalach**-81.85, **thal**-78.03, **oldpeak**-74.79, **slope**-55.34, **exang-**53.75, **chol**-49.61, **age**- 42.07, **sex**-41.63, **trestbps**-38.05, **restecg**- 37.76, **fbs**- 0.002

Ca being the most important and fbs being the least important among all the predictors.

14. **Identify a base-line predictive model in each of the classes of analyses (one for the regression dataset and one for the classification dataset). Explain the rationale for your choice of the baseline models.**

I consider the logistic regression fit to be the baseline model because it is simple and requires less time to run. We will have two options: either we run the same model, or we export the model coefficients and use the corresponding mathematical expression in another place inside the client application. This especially makes deployment of the model outside of python contexts simpler and more natural.

15. **Compute variable importance, where possible, using appropriate metrics of importance. Aggregate the ranking across the models and state the overall (aggregated across models) ranking of each of the predictors.**

Ranking order of predictors according to aggregate variable importance of all models:

Ca-92.05, **cp**-90.134, **thalach**-81.85, **thal**-78.03, **oldpeak**-74.79, **slope**-55.34, **exang**-53.75, **chol**-49.61, **age**- 42.07, **sex**-41.63, **trestbps**-38.05, **restecg**- 37.76, **fbs**- 0.00

3 Conclusions

Based on the results of your predictive models, identify the key implications of your findings for a decision-maker. Be sure to consider the results from each model and the aggregate ranking of the variables in your description of key implications. Considering the different models' fit and their results, which models' results are interpretable? Which models' results are not interpretable? Use the interpretability definition as in the Rudin (2019) paper discussed in the week 10 discussion forum.

Based on the aggregate ranking variable importance, the significant parameters that affect the prediction of cardiac disease are Fluoroscopy results, type of chest pain, maximum heart rate achieved, defect type, and patient's depression level. And the parameters that play the least role in predicting cardiac disease are fasting blood sugar, resting ECG results, resting blood pressure, and sex.

4 Lessons learned

What lessons have you learned as part of building predictive models and interpreting them? Focus on what decisions to manage the time and other resources needed to complete the project. Also explain what insights you have learned that are related to

(a) the relative strengths and weaknesses of the modeling approaches (draw upon the numerical and graphical comparisons of model performance);

Using the CARET library for performing models made code look simple and easy to understand rather than long lines of home-grown code which takes more time. Although the extension of LDA is the QDA model, the logistic fit model accuracy and LDA accuracy were similar with very less difference, but the QDA ranked least among all the models for this dataset. The random forest model with the highest accuracy took a longer time to run than other models.

(b) the applicability of various pre-processing steps

Scaling and centering led to better results, also performing PCA could reduce the run time, but in this dataset, the AIC difference between the original model and the PCA model was dingly small, so I decided to go with the original model.

(c) the use of "home-grown" code (e.g., the for-loop based code for performing K-fold CV and for identifying optimal combination of model-tuning parameters) vs. a framework such as a caret

Frame work with caret seems to be easy to use and understand than the home grown code which is lengthy.

According to the results of this data set, the most efficient Machine learning algorithm in predicting cardiac diseases is the Random Forest model with the highest predictive accuracy rate.