Diameter narrowing or not?

Janani Sundaresan

***Abstract*—This paper analyses the 'Heart Disease Data Set' from UCI Machine Learning Repository [1] and finds the important features through various feature selection methods and build a predictive model to predict if the diameter of the heart is narrowing or not which would lead to heart failure. The final conclusion summarizes all the sections and talks about the future scope and the enhancements that can be made to improve the predictive model that was built.**

I. INTRODUCTION

The dataset consists of 14 attributes out of the total 76 attributes. "The "num" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. The predictive models built concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

II. DATA PREPROCESSING

The data was initially checked for missing values and special characters. The observations with value "? were removed since there were only 6 observations. The categorical variables were already label encoded. The target variable 'num' had more than half of the observations under '0' and rest distributed among 1,2,3 and 4 categories. Before preprocessing, '0' category had 54% of the data and other 4 categories together had 46% thereby indicating an imbalance in data for each category. For simplification purposes, values other than '0' were changed to '1' and the data was almost equally distributed among the two target categories.

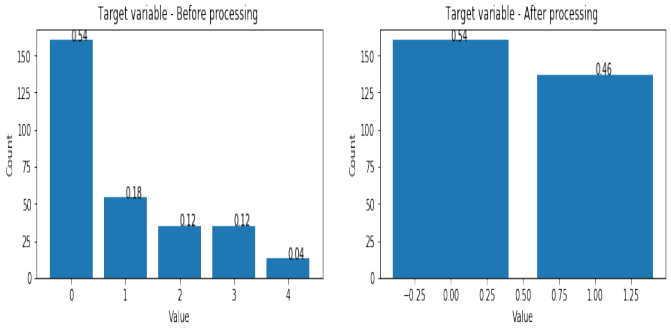

Figure 1: Target variable Distribution

III. FEATURE AND MODEL SELECTION

Usually, principal component analysis is used to reduce the number of features when the data has higher number of dimensions. Since this dataset is relatively small with less dimensions and also most of the correlation coefficients values were smaller than 0.3, PCA was not used. Three different feature selection techniques were employed to identify the best features to predict the target variable – AUC ROC, PPSCORE and Correlation.

"The predictive power score is an asymmetric, data-type-agnostic score that can detect linear or non-linear relationships between two columns. The score ranges from 0 (no predictive power) to 1 (perfect predictive power)." [2] Higher the score better would be the prediction. The PPS matrix was plotted and the row for 'num' variables gives the best univariate predictor for the target. The variables with PPS higher than 0.5 were selected as features for prediction. The 'cp', 'ca' and 'thal' had PPS value higher than 0.58 indicating that they are strong predictors of heart failure.

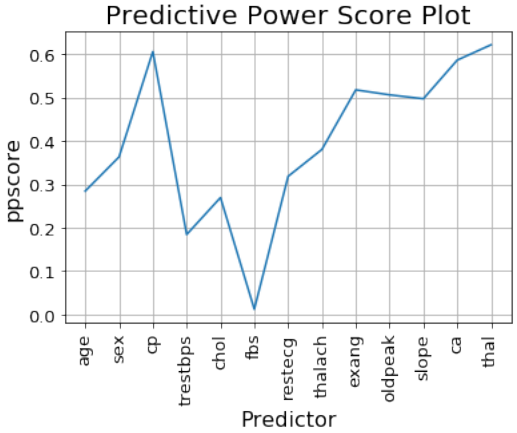

Figure 2: Predict Power Score Plot

Next, AUC ROC curve was used as feature selection technique since this is a binary classification model. The auc roc curve was calculated by adding one variable at a time to the model and only the variables when added to the existing list of variables that improved the model auc roc curve score was finalized for prediction. This technique selected – ''age', 'sex', 'cp', 'chol', 'exang', 'oldpeak'.

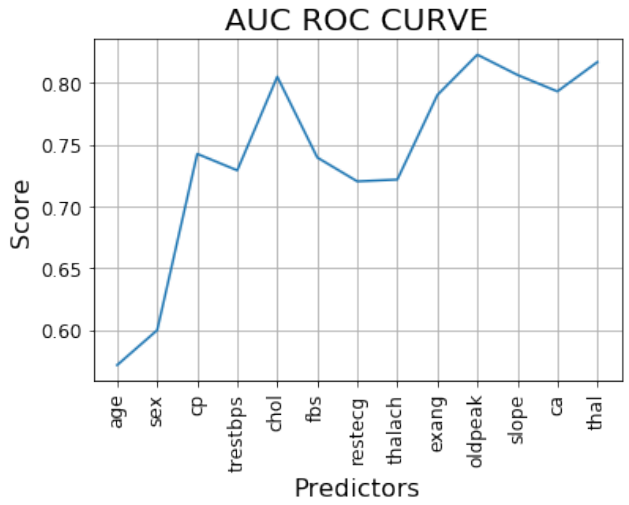

Figure 3: AUC ROC Curve Plot

Next, correlation matrix was plotted and the variables with correlation coefficient value greater than 0.4 with the target variables were selected as features. The selected features are 'cp', 'exang', 'oldpeak'.

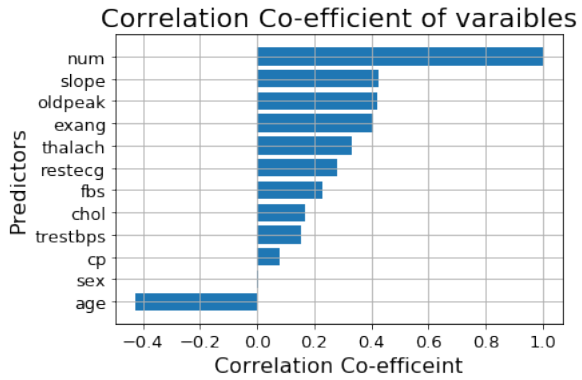

Figure 4: Correlation Coefficient Chart

The features selected from each technique were evaluated against base models of logistic regression and ensemble methods. It can be inferred from TABLE I below that the Logistic Regression model performs better than Random Forest and XGBoost models and the predictive power score feature selection technique gives the best accuracy for the base models. The feature that contributed the most in predicting accurately are 'cp', 'exang', 'oldpeak', 'ca', 'thal'.

Model	Correlation	PPS	AUC
Random Forest	.700	.756	.700
XGBoost	.678	.778	.744
Logistic Regression	.689	.822	.733

Table I – Accuracy for Different Models vs Feature

IV. HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization was performed using grid search cross validation technique, to find the best set of parameters and the best fit was chosen as the optimal model. The predictive power score features were used since it had the best accuracy among all other feature selection techniques. The three base models were tested for a range of parameters corresponding to each model. In case of xgboost and random forest the parameters were same since they both belong to ensemble model. The parameter ranges were carefully chosen to avoid overfitting. 'Max_depth' was considered to be 40, 50 and 'max_features' for all values in range between 1 to maximum number of features. 'min_samples_leaf' were considered for 5 to 15 with a step size of 5 and n-estimators with values 30, 40. The model performance was also tested 'max_lead_nodes' for 2, 3, and 4 values.

V. TRAIN MODEL

Hyperparameter optimization using GridSearchCV returned 'XGBoost' as the best estimator model and that model was used to predict on the test set. The accuracy of the model was 0.855. Most important feature is 'thal' and the best parameter values –

{ 'max_depth': 2,
'max_features': 1,
'max_leaf_nodes': 2,
'min_samples_leaf': 5,
'min_samples_split': 5 }

VI. CONCLUSION

This report presents different models and feature selection techniques that were used to predict heart failure. The same model was tested by keeping all the 5 categories and the accuracy dropped to 0.577 thereby indicating the need for more observations and features. Also, we have different deep learning models that would perform better than the traditional classification model with more data, thereby enabling to build a better prediction model by carefully considering a better feature set.

REFERENCES

[1] Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

[2] https://github.com/8080labs/ppscore