Micah Wiesner
Final - Scripting for Data Analysis

Abstract

Coronary heart disease (CHD) is the leading cause of death and lost life expectancy in the United States population[1]. The economic cost of CHD in the US each year exceeds $80 billion[2]. Utilizing supervised and unsupervised machine learning I examine the prediction of coronary heart disease, over a ten year period, from a variety of risk factors which are in general easy to access with the goal of predicting coronary heart disease accurately.

Two supervised machine learning algorithms were used, a decision tree classifier and random forest classifier, as well as an unsupervised machine learning technique, principal component analysis (PCA), to predict the likelihood of coronary heart disease. The hyperparameters of the decision tree classifier and random forest classifier were tuned to optimize a balance of sensitivity and specificity utilizing three metrics, f1 score, f-beta score, and auc score. Utilizing minority oversampled data the impacts of the imbalanced original dataset were ameliorated. The random forest classifier in conjunction with the minority oversampled data yielded a model with >0.9 specificity and >0.9 sensitivity. PCA was completed on normalized data and log transformed normalized data but unable to clearly distinguish patient groups based on coronary heart disease. This indicates that the variance from each explanatory variable may be similar between patients with and without coronary heart disease.

1.1 Background

Decision tree classification has been used for a variety of domain-specific problems such as predicting hypertension[3], metal contamination[4], formation of tropical cyclones[5], and many other problems. The application within the medical field alone is widespread. Further, random forest classification is especially useful in the medical field in breast cancer research[6], type 2 diabetes[7], and other applications. Principal Component Analysis (PCA) attempts to reduce the dimensionality of a dataset by reducing the dataset into linear combinations of initial variables[8]. By using PCA a dataset can be reduced into the few variables that explain the majority of variance within the dataset. Implementing PCA involves normalizing a dataset so that each variable has a mean of 0 and some variance and then calculating a correlation matrix for all explanatory variables

Our dataset can be classified as an imbalanced dataset, as the majority of patients in our dataset (~85%) are predominantly patients that do not have coronary heart disease while a minority (~15%) actually had coronary heart disease. Imbalanced datasets are common to many areas of machine learning, medical research and fraud analytics are two great examples, and must be addressed in a unique manner. Traditional measures of accuracy, all positive cases divided by all cases, implemented

in our case would be biased by a large case of true negative predictions. There are various approaches to deal with imbalanced data such as oversampling the minority cases, undersampling majority cases, or other more sophisticated techniques such as SMOTE[9].

1.2 Dataset

The dataset was 4,238 x 16. Some of the explanatory variables were gender, age, education, number of cigarettes smoked per day, if the patient was on blood pressure medication, if the patient had diabetes, cholesterol level, and a few others. The explanatory variables were both binary, in the case of gender or whether the patient was on blood pressure medication, and continuous, in the case of the number of cigarettes smoked per day or cholesterol level.

Initial data munging was accomplished by deleting all instances of data, n = 2, which had missing target value, coronary heart disease. All numeric columns with missing values were replaced with the mean of the column, excluding blood pressure data which was missing ~50% of values. This column was excluded from analysis as reporting the mean for such a large number of missing values could create biased data. Systolic blood pressure and diastolic blood pressure were not excluded from the analysis. The target variable, coronary heart disease, was imbalanced as ~85% of patients in the dataset did not have coronary heart disease.

1.3 Exploratory Data Analysis

The average patient in our dataset was age 50, with total cholesterol of 237, systolic blood pressure of 132, glucose levels of 82, and heart rate of 76. The majority of continuous data was normally distributed (figure 1), with the exception of the number of cigarettes smoked per day which was clearly bimodal.

Micah Wiesner
Final - Scripting for Data Analysis

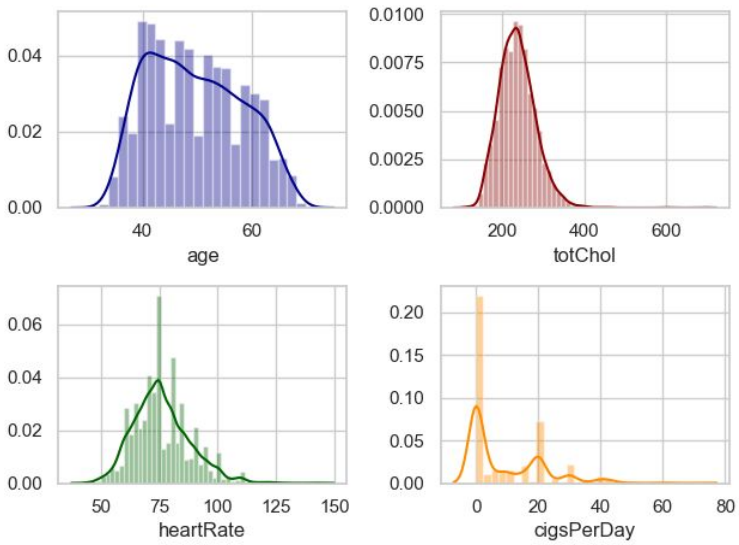

Figure 1: The majority of continuous variables were normally distributed with the exception of cigarettes per day which appeared bimodal.

The dataset was grouped by age group and compared against explanatory variables to understand factors which co-varied with age. Mean age was shown to be covariant with multiple risk factors within our dataset, total cholesterol, BMI, Glucose, and the probability of a patient having coronary heart disease. This led to the belief that it is likely age will be a split within our decision tree and random forest classifier.

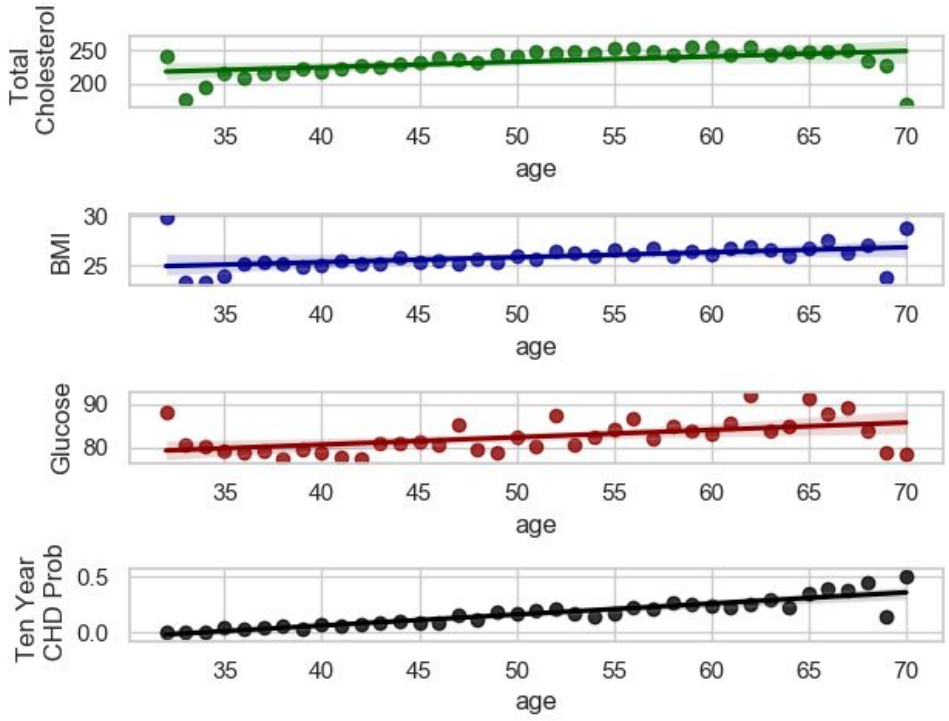

Figure 2: Many risk factors associated with coronary heart disease are strongly correlated to age. In our dataset when patients were grouped by age they showed correlation to cholesterol, body mass index, and glucose levels.

2. Machine Learning Methods

For each model 10-fold cross-validation was utilized to split data between training and testing sets.

2.1 Decision Tree & Tuning

The decision-tree model is a binary model which uses explanatory variables to split unrelated data apart to eventually predict a binary dependent variable. Decision-trees are easy to build and interpret, and can handle numerical and categorical variables. The decision tree has some major advantages over linear regression and logistic regression. Linear regression is not applicable to predict categorical target data. As with linear regression, nonlinear and logistic regressions are not well suited for categorical variables. Decision-trees can identify the most explanatory of variables, which are variables that create the first splits in a decision tree. Decision tree classifiers are also much more easily understandable than artificial neural networks.

Not all decision tree parameters were changed during tuning of the hyperparameters. Using *gridsearchCV*, “entropy” and “gini” criterion were tried, while the maximum depth of the tree was tuned from 1 to 9, and the minimum samples the nodes split into, min_samples_leaf, was tuned from 1 to 8.

A traditional machine learning model can be scored using a traditional metric such as overall accuracy which considers the accuracy of all true cases over all possible cases. With an imbalance dataset the best metric for a model may be AUC, f1-score, or f-beta score. Each of these metrics balances the optimization of sensitivity, also called the true positive rate (eq. 1), against specificity (eq. 2). Even utilizing the proper score can still lead to results that are not very satisfying (table 1) as the number of true negatives outweighs the dataset.

$$1.\ Sensitivity = TP/(TP + FN)$$

$$2.\ Specificity = TN/(TN + FP)$$

Thus for analysis it was necessary to oversample the minority group of patients which had coronary heart disease. Although there are more sophisticated ways to do this, such as SMOTE (Chawla et al., 2002). To do this a synthetic dataset was created containing the normal dataset for all patients without heart disease combined with each instance of patients with coronary heart disease sampled two and three times respectively.

Even using different metrics for scoring the decision tree most models converged on using the ‘entropy’ criterion, a tree with maximum depth of 3, and the minimum number of sample leaves being 6.

2.2 Random Forest & Parameters

The random forest was utilized with the original dataset and two oversampled minority datasets. The random forest chosen had the hyperparameters of 'max_depth': 20, 'max_features': None, 'n_estimators': 15. Random forests utilizing these hyperparameters and utilizing a dataset with twice oversampled minority features and the f-beta scorer created the best model in regards to balancing sensitivity and specificity.

2.3 - PCA

PCA was implemented after normalizing data, and after log transforming and normalizing data. In each case only continuous data variables were used within the PCA.

3. Model Evaluation

3.1 Decision Tree Model Evaluation

Even whilst utilizing the proper metrics the decision tree classifier did not perform well in distinguishing patients that actually had coronary heart disease while using the original dataset (table 1, original). In each of these models run there were a large number of false negatives compared to true positives giving extremely low sensitivity measurements. Each synthetic dataset was created by simply double counting minority instances (synthetic 1) or triple counting them (synthetic 2). Increasing the number of patients that actually had coronary heart disease in our dataset drastically increased the models sensitivity with a small reduction in specificity.

Data	Test Type	Sensitivity	Specificity
Original	F1	0.0566	0.9967
Original	F-Beta	0.0566	0.9967
Original	AUC	0.0503	0.9822
Synthetic 1	AUC	0.5426	0.9679
Synthetic 2	AUC	0.6079	0.9088

Table 1: To account for an imbalance dataset, patients that had tenyearchd were oversampled in our dataset creating a partially synthetic dataset. In analyzing the best performing decision tree three metrics were used f1 score, f-beta score, and roc-auc. All metrics drastically improved through the use of

oversampling of majority cases. The best performing models utilized twice oversampled minority data. For the sake of brevity not all metrics have been shown in the preceding table.

3.2 Random Forest Model Evaluation

For each data set, the original and two oversampled synthetic datasets, random forests outperformed decision trees. Each scoring metric produced sensitivity and scoring results that seemed to be within error rates of one another for a given dataset. The largest increase in model sensitivity occurred from oversampling minority data once over.

Data	Test Type	Sensitivity	Specificity
Original	F1	0.0723	0.9832
Original	F-Beta	0.1092	0.9831
Original	AUC	0.1043	0.9833
Synthetic 1	AUC	0.7300	0.9457
Synthetic 2	F1	0.9341	0.9205

Table 2: To account for an imbalance dataset, patients that had tenyearchd were oversampled in our dataset creating a partially synthetic dataset. In analyzing the best performing confusion matrix three metrics were used f1 score, f-beta score, and roc-auc. All metrics drastically improved through the use of oversampling of majority cases. The best performing models utilized twice oversampled minority data. For the sake of brevity not all metrics are summarized for each model.

3.3 PCA Analysis

PCA analysis was used to understand which variables contain the most variance and how ‘reducible’ the dataset is. To run PCA the continuous variables were used and categorical/binary columns were ignored. After normalizing and reducing it appears that PCA does not clearly define separate clusters based on having coronary heart disease or not.

Micah Wiesner
Final - Scripting for Data Analysis

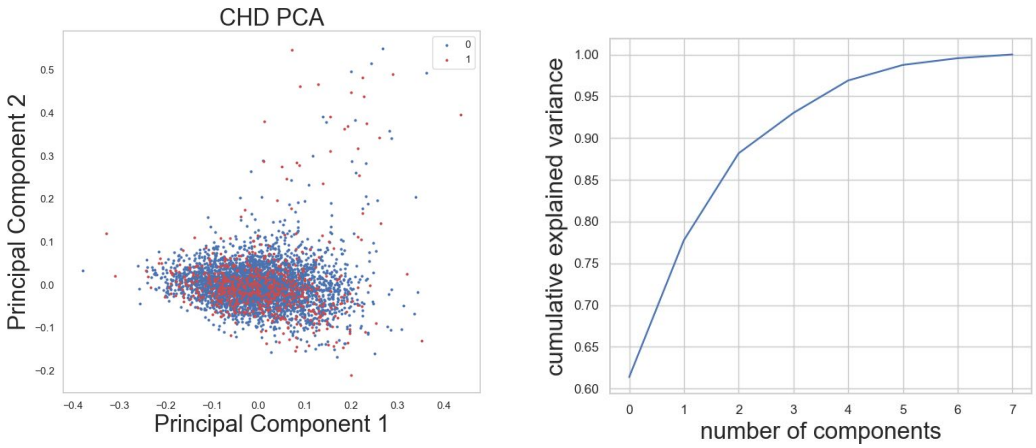

Figure 3: (Left) PCA analysis was unable to separate the groups of interest, patients with coronary heart disease and those without, into clearly clustered groups after reducing the dataset. (Right) The majority of variance in the data can, 88%, be explained by just two components in the data. This could mean that the sources of variances are similar in each group.

The dataset was transformed logarithmically and run with PCA a second time. Log transformed PCA was also unable to demarcate clearly a difference between each group thus suggesting that the variance within each group may be similarly distributed.

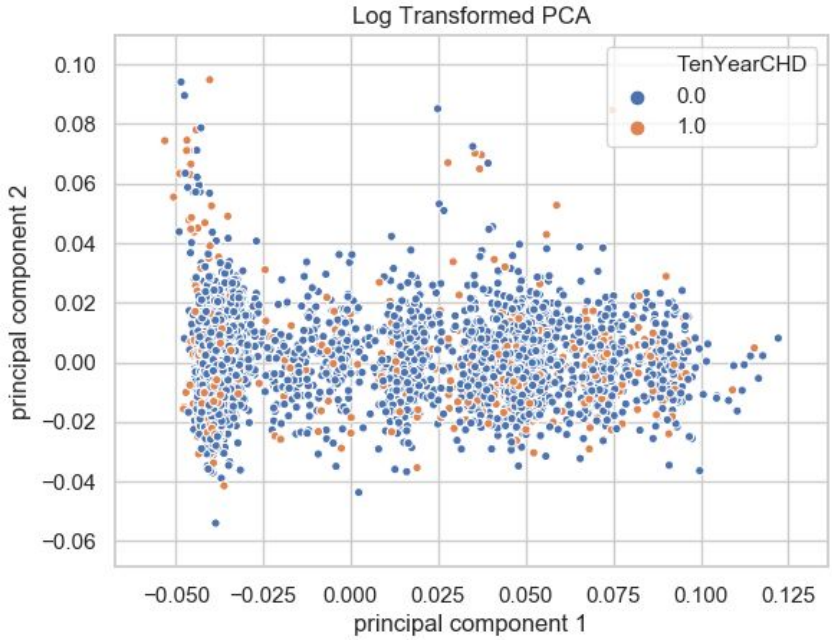

Figure 4: Log transformed PCA did not clearly demarcate a difference between groups of patients that had coronary heart disease and those that did not.

Micah Wiesner
Final - Scripting for Data Analysis

Conclusions

Multiple predictive machine learning models were created to classify patients as being likely to have coronary heart disease within ten years or not based on continuous and categorical explanatory variables. Two supervised machine learning algorithms were utilized in this study, the decision tree classifier as well as random forest ensemble classifier as well as unsupervised machine learning, PCA. Each classifier was run on the original dataset for hyperparameter tuning and subsequently hyperparameters were chosen to be run on datasets with oversampled minority cases.

Due to the imbalance data all models had high specificity rates. Oversampling of minority cases was necessary to create a model with reasonable sensitivity. Oversampling of minority cases once over increased sensitivity from ~0.05 to 0.50 for decision tree classifiers and from ~0.1 to 0.7 for random forest classifiers. The random forest model used over a twice oversampled minority dataset was found to be the most accurate model with respect to specificity and sensitivity and yielded specificity and sensitivity greater than 0.9. Principal component analysis was unable to demonstrate that the groups of patients with and without coronary heart disease had significantly different sources of variation from explanatory variables.

Future work could look at the difference in random forest model performance over minority oversampled data and majority undersampled data. Further, it would be interesting to utilize other transformations in attempting to understand if principal component analysis could be useful in analyzing this dataset. Finally, as more data becomes available it will be necessary to apply our best performing model, random forest classifier, with new testing data to test the sensitivity and specificity from patients in other geographic areas.

Micah Wiesner
Final - Scripting for Data Analysis

1. Karen T. Hicklin, Julie S. Ivy, James R. Wilson, Fay Cobb Payton, Meera Viswanathan, Evan R. Myers. 2018. Simulation model of the relationship between cesarean section rates and labor duration. *Health Care Management Science.* Vol. 25.

2. M. Leshno, U. Goldbourt, I. Pinchuk, D. Lichtenberg. 2018. The cardiovascular benefits of indiscriminate supplementation of omega-3 fatty acids; meta-analysis and decision-making approach. *International Journal of Food Sciences and Nutrition.* 69:5, 549-556.

3. Tayefi, M., Esmaeili, H., Saveri Karimian, M., Amirabadi Zadeh, A., Ebrahimi, M. & Safarian, M. et al. (2017). The application of a decision tree to establish the parameters associated with hypertension. *Computer Methods And Programs In Biomedicine*, *139*, 83-91.

4. Bou Kheir, R., Greve, M., Abdallah, C., & Dalgaard, T. (2010). Spatial soil zinc content distribution from terrain parameters: A GIS-based decision-tree model in Lebanon. *Environmental Pollution*, *158*(2), 520-528.

5. Li, W., Yang, C., & Sun, D. (2009). Mining geophysical parameters through decision-tree analysis to determine correlation with tropical cyclone development. *Computers & Geosciences*, *35*(2), 309-316.

6. Nguyen, C., Wang, Y., & Nguyen, H. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal Of Biomedical Science And Engineering*, *06*(05), 551-560.

7. Xu, W., Zhang, J., & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model.

8. Wold, S., Esbensen, K., Geladi, P. (1987) Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*. Vol., 2. pp. 37-52.

9. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research Vol. 16