Modeling of PIMA Indians Diabetes data

Santanu Mukherjee, Austin Vanderlyn and Brenda Maldonado

August 9, 2022

1 Introduction

The objective of this project is to use one or more data analysis methods that we have learned to analyze whether we can predict if a patient has diabetes based on certain diagnostic measurements included in this dataset. We will use the Pima Indians Diabetes Database, that has 768 observations (rows) and 9 variables (columns). The response variable is Outcome and there are 8 other predictor variables for our analysis.

Our goal is to apply multiple statistical machine learning methods such as Logistic Regression, Bagging, Random Forest, Boosting, SVM(linear), SVM(Radial) and SVM(Radial Tune) to understand model performance based on the same segmentation of 80 percent training data and 20 percent testing data set.

We will be testing multiple models to see which one performs the best for this data and which one preforms the worst. There will be three ways to identify the results; by Accuracy, by Sensitivity and by Specificity. With these measurements we can evaluate to see which model to chose.

Some of the objectives that will be focused on is detailed data exploration to understand the nature of the data, Pre-processing of the data, Modeling the pre-processed data using multiple methods and then predict the onset of the diabetes and concluding the results related to prediction accuracy and errors for all of the models run .

2 Data Structure

2.1 Data Preparation

When it comes to the structure of the dataset, it is in .csv format that has been stored in a data frame.There are several medical predictors that we will be analyzing such as BMI, insulin level, age, glucose concentration, Diabetes pedigree and the one target variable, outcome, which are all in numeric format(int). The measurements are binary and in the form; No Diabetes= 1 and Having Diabetes=0.

As mentioned earlier the data for this project has been split into 80 percent training and 20 percent testing. The data pre-processing prepares the data in having a transformation that either adds or eliminate observations within our dataset.

3 Statistical Learning Method(s)

3.1 Exploratory Data Analysis

We next did some further data exploration on the nature of the dataset. First was checking the correlation between predictors using a correlation. While there appeared to be some slight correlation between Age and Pregnancies, there weren't really any strong correlations to be concerned about.

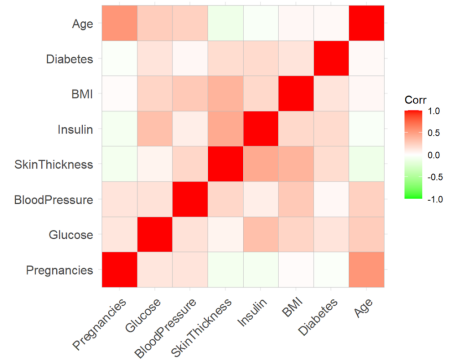

Figure 1: Correlation between Predictors

This means we can expect all of the predictors to be meaningful in predicting the response variable, Outcome.

3.2 Missing Value Imputation

The dataset contains 763 missing values in total, which we then decided to impute using pmm (predictive mean matching). Figure 2 shows the missing values and their distribution among predictors. The vast majority were among Insulin, SkinThickness, and Pregnancies.

All of these missing values were replaced with the mean of that particular predictor, and Figure 3 shows the completion of the pmm process.

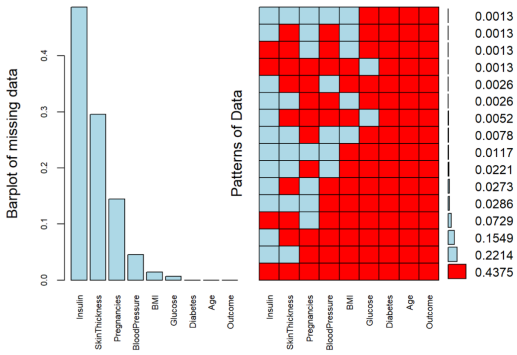

Figure 2: Missing Data and Patterns

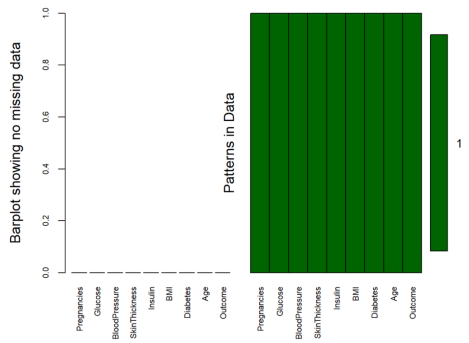

Figure 3: Data Imputation Completed- 0 Missing values

3.3 Data Modeling

Before fitting any of the models, we broke the data up into a training set to fit models on and a testing set used for predictions following a standard 80/20 train/test split.

The first model fitting was a standard logistic regression model, fit using all of the predictors and the glm() function. Once a set of predictions was made with the test set, the initial accuracy rate for the logistic regression model was 0.792.

Next we fit a bagged random forest model with tuning parameters ntrees = 3000 and mtry = 2. The initial accuracy from the bagged model was 0.759, not as good as the basic logistic regression model.

After the bagged random forest model, a standard random forest model with the same tuning parameters as the bagged was fitted, which had an accuracy rate of 0.766. We also ran a boosted random forest with 3000 trees and a shrinkage rate of 0.01, which had an accuracy rate of 0.753.

3.4 ROC Curves

The ROC curves below plot the true positive rate (Sensitivity) in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The Area Under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups.

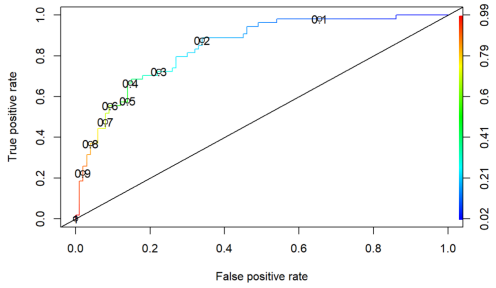

Figure 4: ROC Curve for Logistic Regression

Here are the 3 ROC curves for the tree based models.The accuracy rate for Bagging is 0.759, for Random Forest is 0.766 and for Boosting is 0.753.

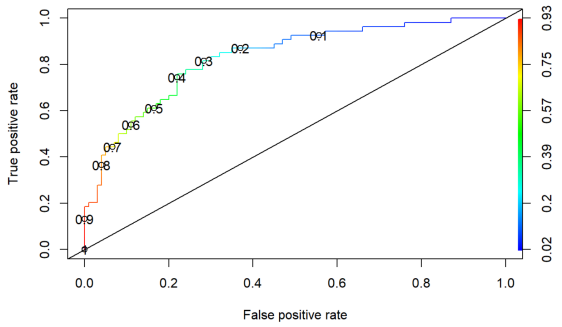

Figure 5: ROC Curve for Bagging

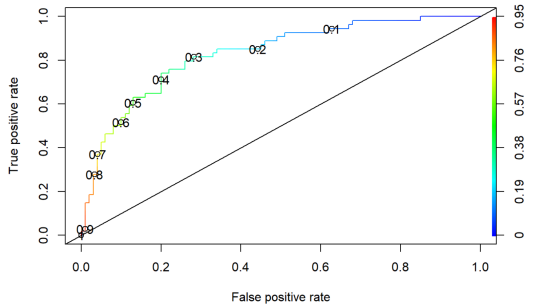

Figure 6: ROC Curve for Random Forest

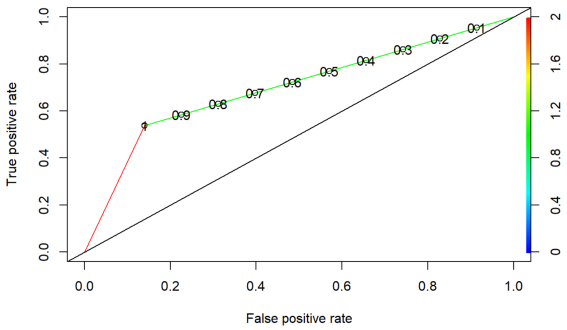

Figure 7: ROC Curve for Boosting

The last three models were all variations on a support vector machine. There was a base SVM model, that returned an accuracy rate of 0.811. Next was an SVM Radial model with a cost of 1 and gamma = 0.1 which returned an accuracy on test predictions of 0.824. Lastly, we fit an SVM Radial model with a grid of cost and gamma tuning parameters that ended up with an accuracy level of 0.694.

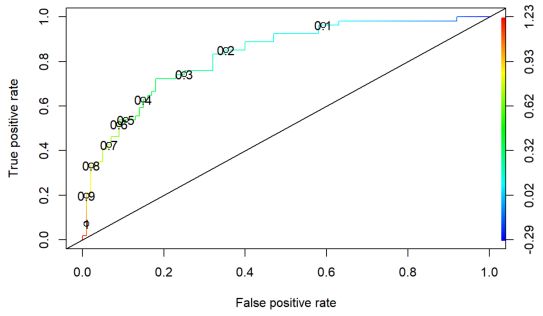

Figure 8: ROC Curve for SVM - Linear

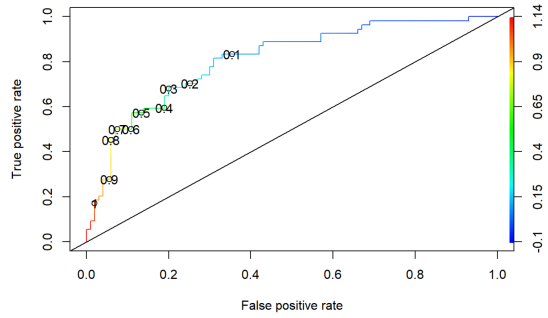

Figure 9: ROC Curve for SVM - Radial

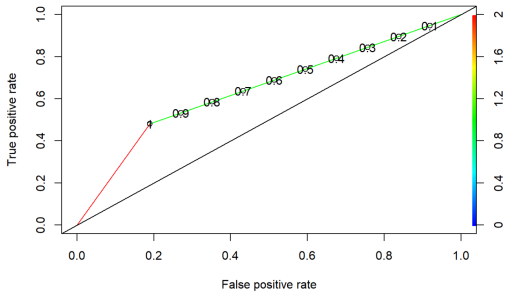

Figure 10: ROC Curve for SVM - Radial Tuning

4 Analysis Results

For all of these models, we fit them using an 80/20 training/testing split. So, we got the best results using the SVM radial, but that doesn't necessarily mean that the SVM radial is the be all end all. You could possibly get better results using a different train/test split or using different patient data.

For instance, while the accuracy on the SVM radial performed best at predicting with the test set, if we look at the ROC curves for each model, the logistic regression model had a pretty decent AUC. So again, with slightly different data or a different train/test split, the logit model could be the best model at predicting patient outcome.

In addition, after the SVM radial, the logit model had the best trade-off of sensitivity and specificity, which shows that it could be equally as good at predictions as the SVM radial with slightly different

data or train/test split.

Table 1: Comparison of Model Statistics

	Accuracy	Sensitivity	Specificity
Logistic Regression	0.792	0.830	0.833
Bagging	0.759	0.789	0.689
Random Forest	0.766	0.796	0.696
Boosting	0.753	0.782	0.682
SVM - Linear	0.811	0.807	0.800
SVM - Radial	0.824	0.845	0.841
SVM- Radial Tune	0.694	0.726	0.594

5 Conclusions, Discussions, and Bibliography

Determining what is the "best" and "worst" model can be a difficult question to answer, but based upon the predictions we have here, it appears that both the best and worst performing models were variations on support vector machines. The worst performing model was the SVM radial tune, with an accuracy of classification of 69.4 percent, and the best performing model was the SVM radial without tuning parameters, that had a classification accuracy of 82.4 percent.

However, as discussed in the analysis section, several of the other models, most notably the logistic regression model, could perform equally as well as the SVM radial at real-world predictions.

5.1 References

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?select=diabetes.csv.