BIOS 735 - Project Proposal

Anil Anderson[1], Arthi Hariharan[1], David Hu[1], Yumei Yang[1], and Álvaro Quijano[1]

[1]Department of Biostatistics, UNC Chapel Hill

1 Introduction

We are interested in examining the risk factors for diabetes apart from blood glucose level which can be identified from blood tests. We will be using the Health Test by Blood dataset from Kaggle (Anjali, n.d.). This is a cross-sectional data on clinical variables related to cardiovascular and kidney function. There are 5132 observations and 10 variables: Age, Gender, BMI, Total Cholesterol, Triglycerides, High-Density Lipoprotein, Low-Density Lipoprotein, Creatinine, Blood Urea Nitrogen, and Diabetes status.

In this project, we propose to build a classification model using penalized logistic regression and Random Forest for predicting the diabetes status of patients based on blood test results. A penalized logistic regression with a regularization term such as Lasso or Ridge penalty will help mitigate problems arising from overfitting and multicollinearity in the variables. Employing Random Forest can help identify patterns in the data using importance metrics to identify the variables which predict diabetes status.

2 Aims

- To identify and assess the risk factors associated with Diabetes, including Body Mass Index (BMI), total cholesterol (Chol), triglyceride levels (TG), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), creatinine (Cr), and Blood Urea Nitrogen (BUN), while controlling for age and gender.
- To evaluate and compare the performance of the Penalized Logistic Regression and Random Forest methods with respect to model fitting results.
- To develop a predictive model for diagnosing diabetes in future patients, using various clinical and demographic factors.

3 Methods

3.1 Penalized Logistic Regression

Using logistic regression without additional regularization might yield favorable results given our 9 covariates and 5132 observations. However, to mitigate the risk of over-fitting for diabetes prediction and select the most relevant risk factors, we opted for penalized logistic regression. This approach reduces the chances of over-fitting while controlling the complexity in the model.

Let x_{ij} denote the observed data, where i corresponds to patients and j corresponds to the independent variables such as age, gender, Body Mass Index (BMI), total cholesterol (Chol), triglyceride levels (TG), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), creatinine (Cr), and Blood Urea Nitrogen (BUN), with $i = 1, \ldots, n$, with n = 5132 and $j = 1, \ldots, 9$. Let y_i represent the binary outcome for observation i, indicating the presence or absence of diabetes. Our goal is to classify observations based on the binary outcome y_i, considering the covariates represented by X and identifying the risk factors associated to diabetes.

The log-likelihood for the model can be written as,

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \{y_i \log(\pi(x_i;\beta))\} + (1-y_i)\log(1-\pi(x_i,\beta))\} && (1) \\ &= \sum_{i=1}^{n} \{y_i \beta^T x_i + \log(1+e^{\beta^T x_i})\} && (2) \end{aligned}$$

And can be penalized (using a quadratic Ridge penalty) in the following way,

$$l(\beta) = \sum_{i=1}^{n} \{y_i \beta^T x_i + \log(1+e^{\beta^T x_i})\} - \frac{\lambda}{2}\sum_{j=1}^{m} \beta_j^2 \tag{3}$$

Where the complexity (Ridge) parameter λ controls the size of the coefficients β_j. Thus, our objective is using general optimization techniques to provide a solution to the parameters β_j and cross-validation to choice of the regularization parameter (Schimek, n.d.).

For penalized logistic regression, we plan to implement the EM algorithm outlined in Module 2 to estimate β_j.

3.2 Random Forest

We will also use a random forest to predict diabetes. Random forests use an ensemble of decision trees to make predictions in classification and regression problems. The algorithm for building these trees involves a combination of "bagging," i.e. sampling with replacement from the original data, and randomly subsetting predictors. We will use the *randomForest* package in R to build the forest.

Whereas the penalized logistic regression model employs shrinkage to identify key risk factors for diabetes, random forests give several importance metrics that identify the variables with the most predictive power. One such metric is accuracy-based importance, which is calculated by averaging the decrease in prediction accuracy across trees when the values of a certain variable in the out-of-bag sample are shuffled and all others are held constant. Another such metric is Gini-based importance. The Gini impurity for binary classification problems is defined as

$$GI = 2p - 2p^2,$$

and the algorithm chooses the variable to split at each node by calculating the decrease in Gini impurity. The average Gini decrease across trees for a certain variable thus quantifies the importance of that variable in making predictions. We will use both metrics to identify important predictors of diabetes.

For the random forest model, we plan to use R functions outlined in Module 3.

4 Analysis Plan

We will report accuracy, true positive rate(TP), true negative rate(TN), precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to compare performances of the two models. We define the metrics as follows:

Accuracy is the ratio of correctly predicted cases (both true positives and true negatives) to the total number of instances, i.e.,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Precision is the ratio of true positive cases to the total number of cases predicted as positive, i.e.,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall is the ratio of true positive cases to the total number of actual positive cases., i.e.,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

And

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

We will compare the two models using a comprehensive assessment of these metrics, as well as discuss any trade-offs between the metrics. We will use a stratified 80:20 train-test split for both models.

References

Anjali, S. (n.d.). *Health test by blood dataset: Blood test for health dataset.* Kaggle. Retrieved from `https://www.kaggle.com/datasets/simaanjali/diabetes-classification-dataset`

Schimek, M. G. (n.d.). Penalized logistic regression in gene expression analysis. Retrieved from `https://www.cs.cmu.edu/ gpekhime/Projects/CSC2515/Refs/Regression/schimek.pdf`