Survival Analysis and Machine Learning Methodologies for Bone Marrow Cancer

Yitian Cai, Rudra Guin, Jorge Portugal

12/17/2021

Introduction

For this project, our group chose to perform statistical analysis using the Clinical Data Set from the Clinical Package of bone marrow cancer data. It was collected from the National Cancer Institute (NCI) genomic data portal. The version of the dataset we used was released by October 29, 2021. The bone marrow cancer can damage the bones, immune system, kidneys, and red blood cell count. The dataset contains 16,029 different cases of bone marrow cancer. We planned on using it for survival analysis and machine learning methodologies.

Data Description

The dataset contains 47 clinical variables. At the beginning, we tried to include the biospecimen data in order to increase the dimension of our dataset, but we found that most of those genomics data were missing, which would not provide us too much help. So, we decided to stick with the clinical data of bone marrow. Vital status means whether the patients were alive or dead at the time the data were collected. Our dataset contains 9,036 alive cases and 3,551 dead cases. Others were either missing or recorded as “unknown” or “not reported”.

Data Preprocessing

The initial dataset has several missing values among the different features. So, before going into analysis and modeling, we first clean up and pre-process the data. First, we remove all the observations that have missing values for vital status, because vital status is the one we wish to forecast or predict. Then, we decided to remove variables with more than 8,000 missing values. Because it means that more than half of the observations in these variables were missing, which we believe will cause too much inaccuracy after data imputation. After that, we used mice() to impute those missing values. The mice package implements a method to deal with missing data and it generates Multivariate Imputations by Chained Equations. And finally, we randomly split the dataset into two where the training set contains 70% of the data, and the test set contains the rest 30% of the data.

Analysis

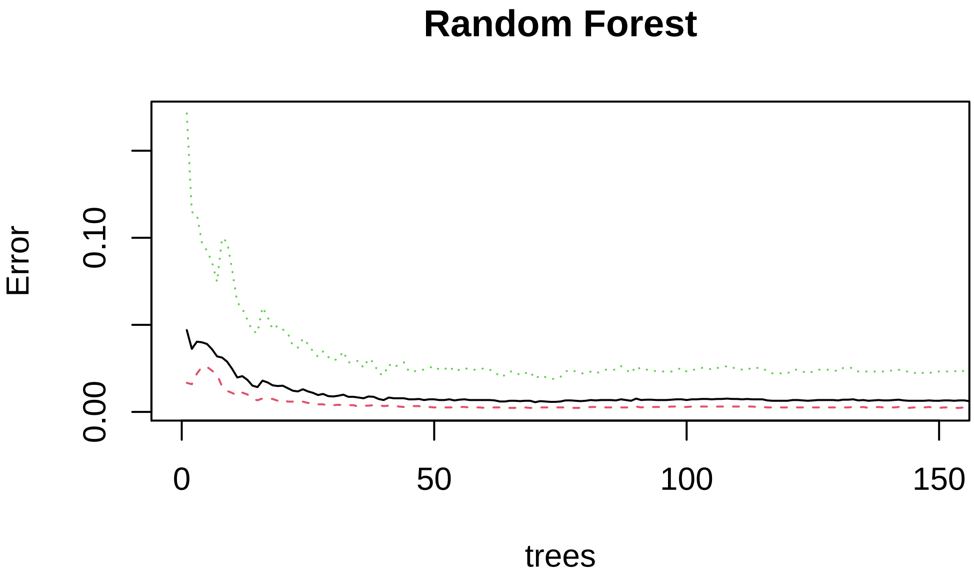
Survival Analysis

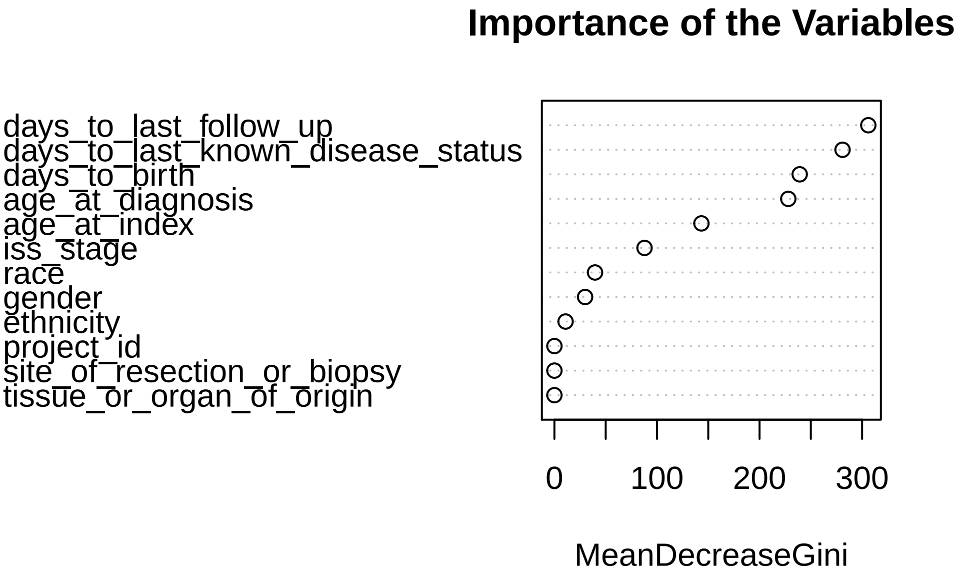
We first conducted some survival analysis

Machine Learning Modeling and Analysis

We also performed some machine learning modeling and evaluated the effectiveness of the models’ prediction power of the vital status of certain patients.

Random Forest





```
## [1] "Training set confusion matrix and accuracies:"
```

```
##           Reference
## Prediction Alive Dead
##      Alive  3896    2
##      Dead      0  947
```

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##      0.9995872      0.9986885      0.9985096      0.9999500      0.8041280
## AccuracyPValue  McnemarPValue
##      0.0000000      0.4795001
```

```
## [1] "Testing set confusion matrix and accuracies:"
```

```
##           Reference
## Prediction Alive Dead
##      Alive  1656    1
##      Dead      4  410
```

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   9.975857e-01   9.924320e-01   9.943749e-01   9.992156e-01   8.015451e-01
## AccuracyPValue  McnemarPValue
##  3.216699e-188   3.710934e-01
```

Neural Network

After making dummy variables for every categorical feature and normalizing each numerical feature, several instances with missing values were removed, therefore the total number of instances in the overall neural network dataset was 6916, before it was split into training and testing subsets. Prior to the removal of instances with missing values for the numerical features, there were a total of 12587 instances in the dataset.

It takes under 50 minutes to generate the neural network model, trained using the training subset. It could probably be sped up using services such as Great Lakes, etc.

```
## [1] "Training set confusion matrix and accuracies:"

##           Reference
## Prediction    0    1
##          0  859   16
##          1   88 3878

##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   9.785168e-01   9.297138e-01   9.740290e-01   9.824138e-01   8.043793e-01
## AccuracyPValue  McnemarPValue
##  1.789780e-305   3.351831e-12

## [1] "Testing set confusion matrix and accuracies:"

##           Reference
## Prediction    0    1
##          0  344   15
##          1   69 1647

##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   9.595181e-01   8.664742e-01   9.501238e-01   9.675839e-01   8.009639e-01
## AccuracyPValue  McnemarPValue
##  4.504814e-100   7.347871e-09
```

Results

Conclusion

Division of Work:

Yitian: worked on initial data preprocessing by transforming all variables into numeric or factor, then worked on data imputation and data cleaning, and finally implemented and evaluated the random forest model.

Rudra: worked on initial data preprocessing by converting empty values to `NA`, then did further preprocessing of the data for the neural network model, and finally implemented and evaluated the neural network model.

Jorge: