Data 630 – Fall 2018

Assignment 2 – Naïve Bayes Analysis of Fracture Risks in Women

Kenneth Lulie

Dr. Gates

University of Maryland University College

October 14, 2018

**Naïve Bayes Analysis of Bone Fracture Risk using the GLOW Dataset**

**Introduction**

Osteoporosis is a medical issue in which bones become brittle and fragile. Osteoporosis literally means “porous bone”. People suffering from Osteoporosis experience lower density or mass in their bones which weakens bone structural integrity. Osteoporosis sufferers may fracture their bones from a simple fall or in extreme cases even from a cough, sneeze, or bumping into furniture. Osteoporosis is caused by an imbalance between the creation of new bone tissue and reabsorption of old bone tissue. Osteoporosis may be caused by not enough new bone growth, too much bone being reabsorbed, or both simultaneously. Low bone growth can be due to insufficient levels of calcium, phosphate, or vitamin D in the patient’s diet. Age is another risk factor as the human body typically peaks in bone density in the early twenties, and then decreases as the body ages. Osteoporosis is called a silent disease as there are no signs indicating lowering bone strength until you experience a fracture.

Osteoporosis can affect anyone, but the groups at highest risk are white and Asian women who have undergone menopause (Mayoclinic, 2016). Risk factors include physical inactivity and a sedentary life style, a high intake of alcohol, prolonged use of corticosteroids, low body weight and recent weight loss, drop in sex hormones due to menopause, family history of Osteoporosis, being over age 50, and smoking. It is estimated that 10 million people in the United States have Osteoporosis with another 44 million people suffering from low bone mass . Of these 54 million people, 68% are women with the remaining 32% men. Studies suggest that one of every two women over the age of 50 will have an osteoporosis- related fracture in their life time, while men over 50 are estimated to have a one in four chance. This disease is not only an issue to the patients and their families, but also to the health care system with an estimated $19 billion dollars spent on Osteoporosis-related bone breaks per year.

There are different options for treatment and prevention of Osteoporosis. Beyond lifestyle changes such as increasing exercise, changing diet, smoking cessation and reduction of alcohol intake, medicines exist that can help prevent bone loss or help rebuild bone resulting in lower chances of bone fractures (NOF, 2018). These medicines include Estrogen Therapy to replace the sex hormones lost when undergoing menopause to medicines that help rebuild bone such as Abaloparatide (Tymlos) and Teriparatide (Forteo) (Mayoclinic, 2016). However, as Osteoporosis is often a silent disease it is essential to understand the risk factors so treatment and prevention can begin before the first occurrence of bone fracture. In this analysis, the impact of different factors such as age, history of prior fractures, BMI, menopause status, family history of Osteoporosis, muscle strength, smoking status, and self-estimated risk of fracture will be tested to identify their impact as risk factors. The specific aim of this analysis is to use the Naïve Bayes algorithm to test the relationships of these different factors and build a model that predicts the risk of a fracture. The model will be able to predict the specific odds for a given patient to experience a fracture in the next five years and show how each factor affects the total risk of a fracture event.

**Analysis and Model Demonstration**

**Subsection: Data Information, Cleaning, and Preprocessing**

The data used in this analysis comes from the Global Longitudinal study of Osteoporosis in Women (GLOW). GLOW was a cohort study involving 723 physicians and 60,393 women patients over 55 years (Glow, 2009). The study focused on following these 60,393 women to improve understanding of international patterns of susceptibility, recognition, management, and outcomes of care in women aged 55 years and older at risk for fragility fractures.

The analysis of this study was implemented using Rstudio version 3.5.1. There were 500 patients in the dataset provided. There were no missing values in the observations provided. The data as provided contained 15 variables including Patient ID, Site ID, Physician ID, if the patient had previously had a fracture, age at enrollment, weight at enrollment, height at enrollment, BMI at enrollment , if the patient experienced menopause before the age of 45, if their mother had a hip fracture, if they need their arms to stand from a chair, if they are a current or former smoker, their self-reported risk of fracture a Fracture Risk score calculated from their medical history and age, and if they experienced a fracture in the first year after enrollment in the study.

Part of the data preprocessing involved the removal of unneeded variables. The patient ID and the physician ID were unique identifiers and were removed from the dataset as unnecessary for the analysis. Additionally, the site ID was removed as no information was provided allowing for identification of the sites. Further, as Naïve Bayes assumes the independent variables are independent from each other, weight and height were removed as their relationship is coded into the BMI variable. For the same reason, the Fracture Risk Score variable was also removed as it is calculated using other variables in the dataset. At this stage, all remaining variables except for Age and BMI were converted from integer as initially provided to factor to meet the requirements of the Naïve Bayes algorithm.

A summary of the frequency of the categorical variables are presented in **Table 1**. There

Table 1. Observation Frequencies in GLOW dataset.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Prior Fracture | Menopause before Age 45 | Mother Had Hip Fracture | Required Arms to Stand from Chair | Current or Former Smoker | Fracture Within 1 Year of Entering Study | Self-Rated Risk of Fracture | | |
| No | 374 | 403 | 435 | 312 | 465 | 375 | Low | Medium | High |
| Yes | 126 | 97 | 65 | 188 | 35 | 125 | 167 | 186 | 147 |

were 375 cases without a fracture within 1 year in the dataset, and 125 cases that did experience a fracture. There were few smokers in the dataset with only 35 cases, but 126 patients had previously experienced a fracture. A descriptive analysis including mean, median, 1st and 3rd quartiles, min and max of the Age and BMI variables are provided in **Table 2**. The boxplot data visualization for Age and BMI show that Age is skewed below the median, and that BMI contains some outliers (**Figure 1**).

The Naïve Bayes algorithm requires that numeric variables be converted to factor before running. Age and BMI were thus converted to factor variables using unsupervised discretization with the equal frequency method into six factors each. Equal frequency distribution was chosen as it handles outliers well. The frequency distribution and factor ranges of Age and BMI are provided in **Table 2**.

Table 2. Distribution of Age and BMI before and after Discretization

|  |  |  |  |
| --- | --- | --- | --- |
| Before Discretization | | | |
| Age | | BMI | |
| Min | 55 | Min | 14.88 |
| 1st Qu. | 61 | 1st Qu. | 23.27 |
| Median | 67 | Median | 26.42 |
| Mean | 68.56 | Mean | 27.55 |
| 3rd Qu. | 76 | 3rd Qu. | 30.79 |
| Max. | 90 | Max. | 49.08 |
|  | | | |
| After Discretization | | | |
| Age | | BMI | |
| [55,59) | 70 | [14.9,22.1) | 84 |
| [59,63) | 94 | [22.1,24.1) | 83 |
| [63,67) | 70 | [24.1,26.4) | 83 |
| [67,72) | 92 | [26.4,29) | 83 |
| [72,78) | 84 | [29,32.9) | 82 |
| [78,90] | 90 | [32.9,49.1] | 85 |

**Subsection: Analysis and Model Methods**

The preprocessed data was used in this analysis with the Naïve Bayes algorithm for classification to predict if patients did or did not experience a fracture in the first year after enrolling in the study. A simultaneous result is also data on how each variable affects the probability of experiencing a fracture during that time period. Naïve Bayes is capable of not only classification, but also showing the odds, or how confident that classification is. A model that can help identify the risk factor of fractures can serve an important tool in determining which patients are at greatest risk as well as what level of treatment and life style adjustments might be necessary as well as giving general guidance to the public as to the impact of life style choices on the risk of developing Osteoporosis.

The Naïve Bayes algorithm is a supervised classification method which assumes that all independent variables in the data are conditionally independent (Han, 2011). It also assumes that all independent variables are equally important when calculating the value of the dependent variable. These limitations are often violated in actual datasets which is why the algorithm is named “Naïve”. Even when these assumptions are violated, Naïve Bayes can still provide strong results off of limited amounts of data. The algorithm produces an output where the probability of classification to a specific class is based on the product of conditional probabilities for each variable value divided by the probability of an arbitrary observation belonging to that specific class based on prior observations.

One of the strengths of Naïve Bayes is that it can be used to calculate probabilities which can be more important for actual applications than the classification itself. An example is that the percent of risk of default is more useful for a loan officer than a simple prediction as to if that person will default. Using this information a loan officer can adjust the loan terms to properly account for the risk of default. Similarly, this analysis of the GLOW dataset with Naïve Bayes could yield similar capabilities to help doctors predict the risk of fractures for their patients.

To measure the accuracy of the Naïve Bayes model a confusion matrix will be used to review the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). From these numbers Error Rate (FP + FN)/Total Number of Predictions), Accuracy (TP + TN/Total Number of Predictions), Sensitivity (TP/ Total Actual Positives), and Specificity (TN/Total Actual Negatives), Precision (TP/TP + FP), False Positive Rate (FN/TN+ FN) can be calculated and used to measure the model.

This analysis reviewed the results of three different models using the Naïve Bayes algorithm. The data was randomly sampled with 70% going to a training dataset, and the remaining 30% going to the testing dataset. Each model was trained on the training set, and then used to predict fractures on the test dataset. The first model used all remaining variables after data preprocessing, specifically if the patient had a prior fracture, the age of the patient, the BMI of the patient, if the patient had experienced menopause before age 45, if the patient’s mother had a hip fracture, if the patient required arms to stand from a chair, if the patient was a current or former smoker, and the patients self-rated risk of fracture were used to predict if the patient would experience a fracture within one year of enrollment. The second model used the same parameters except that the calculated fracture score was re-added into the dataset to review the addition of this calculated score on the accuracy of the model. The Fracture score was discretized using equal frequency with 5 ranges. Third, the fracture score was removed and the Age and BMI variables were run with cluster discretization run using a total of five factors instead of the original six to see if the change would allow for greater accuracy. Each model was evaluated using a Confusion Matrix, as well as the Error Rate, Accuracy, Sensitivity, Specificity, Precision, and False Positive Rate.

**Results**

The first model produced had an accuracy of 74.32% when run against the test data **(Table 3)**. The specificity rate was 91.82% with the false positive rate at 8.18% indicating that the model did a good job of correctly classifying patients who wouldn’t have a fall. However, the precision rate was only 50% indicating that about only one in two patients predicted to have a fracture event did in fact have a fracture event. The sensitivity was also low, with only 23.68% of patients who would experience fractures being classified. The review of the covariates shows that age appeared to be a major factor with patients in the 78 to 90 range having a 3.82 odds higher chance of developing a fracture than people in the 55 to 58 range **(Appendix 1)**. Additionally, having had a prior fracture was another risk factor identified with having approximately twice the odds of developing a fracture than those who did not. Further, requiring arms to get out of a chair was another strong risk factor. A visualization of the confusion matrix is provided in **Figure 2.**

Table 3. Confusion Matrix and Evaluations of Model 1

|  |  |  |
| --- | --- | --- |
| Model 1 - Confusion Matrix | | |
|  | Predicted False | Predicted Positive |
| Actual False | 101 | 9 |
| Actual Positive | 29 | 9 |
| Error Rate | Accuracy | Sensitivity |
| 25.68% | 74.32% | 23.68% |
| Specificity | Precision | False Positive Rate |
| 91.82% | 50.00% | 8.18% |

The second model where fracture score was added had an accuracy of 70.78%, showing a decrease from the first model **(Table 4)**. The model did show an improvement in sensitivity correctly picking out 31.43% of the patients who would experience falls, but showed decreases in all other metrics. This is not a surprising outcome, as Naïve Bayes assumes that the independent variables are independent from each other, and each equally important. As this model reintroduced the fracture score which was calculated from the other variables included in the analysis, the variable violated these rules. A visualization of the confusion matrix is provided in **Figure 3**, and a description of the model produced is provided in **Appendix 2**.

The third model where Age and BMI were discretized in the cluster method had an accuracy of 72.41%, comparable to the first model. Interestingly, the precision and sensitivity were increased over the first model. This indicates that the model did a better job of picking out the patients who would experience fractures, and a better job of only predicting patients who would experience fractures. A visualization of the confusion matrix is provided in **Figure 4**, and a description of the model is provided in **Appendix 3.** This model could be more useful than the first model for certain applications. For example if determining which patients should receive a medicine with strong side affects the highest risk patient population would need to be identified without including lower risk patients. The only patients you would wish to receive the medicine might be only the patients with the highest risk if the side-effects are strong and negative.

**Conclusion**

This analysis showed that the variables provided could be used to build a model to classify patients and review the impact that each variable has on the total risk of experiencing fractures. Specifically, it was found that a prior history of fractures, older ages, and requiring arms to sit up from a chair were strong risk factors, and that BMI, smoking, and having a mother who experienced a hip fracture were weak risk factors. Further, the models developed were only weakly effective, with a highest accuracy of 74.32%.

Analysis of risk factor for Osteoporosis was found to be helpful, and generally assist in identifying the highest patients at risk. Research into Osteoporosis risk factors is essential to understanding Osteoporosis and educating patients, caregivers, and doctors on how different lifestyle and medical backgrounds lead to different outcomes and predicting which patients are at the highest risk and require medical interventions in the form of exercise and medication.

Future studies could improve on the limitations of this analysis. First, the total number of observations included in this dataset was low with only 500 observations. A higher number of observations is needed to have greater confidence in the findings of this analysis. Second, more variables would also assist in developing a more accurate model as there was no information about medicines or lifestyle changes of the patients such as increasing the amount of weight-bearing exercise, changing diet, or taking medication. Additionally, the algorithm was also limited as the variables likely violated the assumptions of the algorithm, namely the requirement for each variable to be independent from each other and that each variable is equally important.

**References**

Han, Kamber, and Pei (2011). Data Mining: Concepts and Techniques, Third Edition Retrieved September 14, 2018 from http://hanj.cs.illinois.edu/cs412/bk3/01.pdf

Osteoporosis. (2016, July 07). Retrieved from https://www.mayoclinic.org/diseases-conditions/osteoporosis/symptoms-causes/syc-20351968

Insights from the Global Longitudinal Study of ... (n.d.). Retrieved from https://www.researchgate.net/publication/261770421\_Insights\_from\_the\_Global\_Longitudinal\_Study\_of\_Osteoporosis\_in\_Women\_GLOW

Learn What Osteoporosis Is and What It's Caused by. (n.d.). Retrieved from https://www.nof.org/patients/what-is-osteoporosis/

Figure 1. Boxplot of Age and BMI in GLOW dataset.

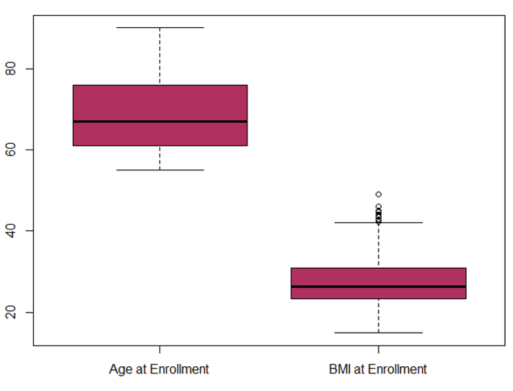
****

Figure 2. Visualization of Model 1 Confusion Matrix

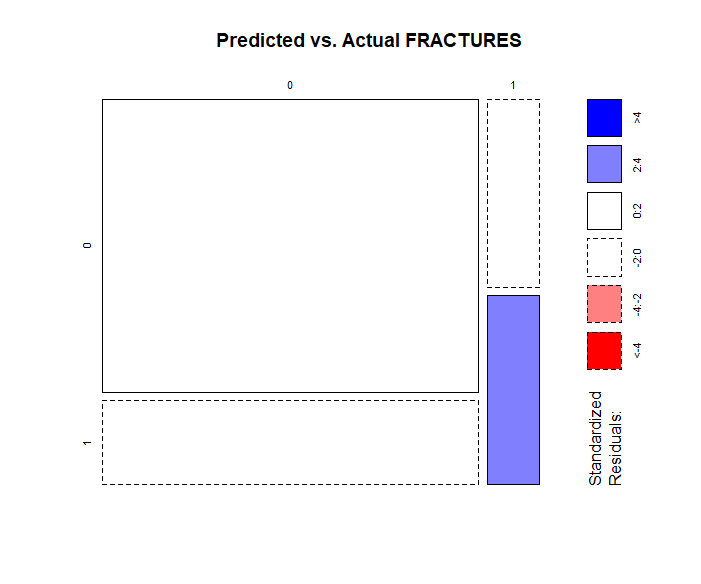


Table 4. Confusion Matrix and Evaluations of Model 2.

|  |  |  |
| --- | --- | --- |
| Model 2 - Confusion Matrix | | |
|  | Predicted False | Predicted Positive |
| Actual False | 98 | 21 |
| Actual Positive | 24 | 11 |
| Error Rate | Accuracy | Sensitivity |
| 29.22% | 70.78% | 31.43% |
| Specificity | Precision | False Positive Rate |
| 89.91% | 34.38% | 17.65% |

Figure 3. Visualization of Model 2 Confusion Matrix

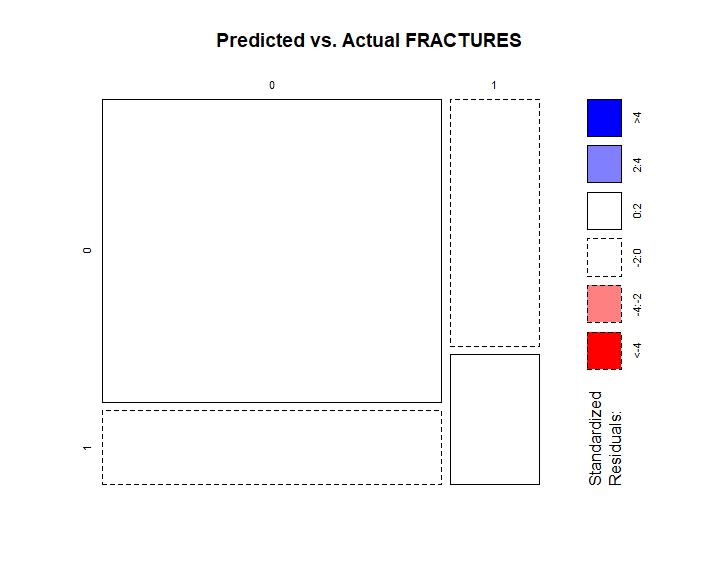
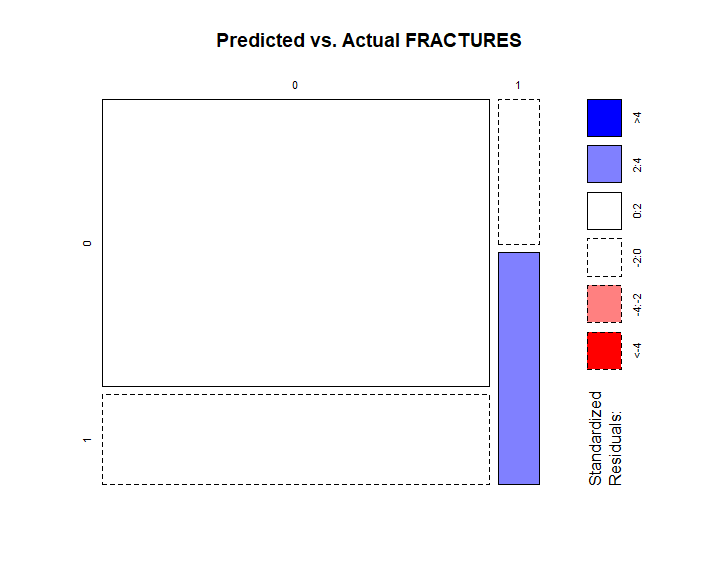


Table 5. Confusion Matrix and Evaluations of Model 3.

|  |  |  |
| --- | --- | --- |
| Model 3 - Confusion Matrix | | |
|  | Predicted False | Predicted Positive |
| Actual False | 93 | 10 |
| Actual Positive | 30 | 12 |
| Error Rate | Accuracy | Sensitivity |
| 27.59% | 72.41% | 28.57% |
| Specificity | Precision | False Positive Rate |
| 88.57% | 54.55% | 9.71% |

Figure 4 Visualization of Model 3 Confusion Matrix



**Appendix 1**

**Model 1 Output**

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

0 1

0.7528409 0.2471591

Conditional probabilities:

PRIORFRAC

Y 0 1

0 0.7962264 0.2037736

1 0.5977011 0.4022989

AGE

Y [55,59) [59,63) [63,67) [67,72) [72,78) [78,90]

0 0.14716981 0.21132075 0.14339623 0.20377358 0.14339623 0.15094340

1 0.09195402 0.12643678 0.13793103 0.10344828 0.19540230 0.34482759

BMI

Y [14.9,22.1) [22.1,24.1) [24.1,26.4) [26.4,29) [29,32.9) [32.9,49.1]

0 0.1584906 0.1773585 0.1811321 0.1547170 0.1471698 0.1811321

1 0.1839080 0.1839080 0.1839080 0.1724138 0.1379310 0.1379310

PREMENO

Y 0 1

0 0.8000000 0.2000000

1 0.7701149 0.2298851

MOMFRAC

Y 0 1

0 0.8981132 0.1018868

1 0.8160920 0.1839080

ARMASSIST

Y 0 1

0 0.6490566 0.3509434

1 0.4597701 0.5402299

SMOKE

Y 0 1

0 0.92452830 0.07547170

1 0.93103448 0.06896552

RATERISK

Y 1 2 3

0 0.3471698 0.3811321 0.2716981

1 0.2298851 0.4137931 0.3563218

**Appendix 2**

**Model 2 Output**

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

0 1

0.7398844 0.2601156

Conditional probabilities:

PRIORFRAC

Y 0 1

0 0.8085938 0.1914062

1 0.5444444 0.4555556

AGE

Y [55,59) [59,63) [63,67) [67,72) [72,78) [78,90]

0 0.1484375 0.1992188 0.1406250 0.2148438 0.1640625 0.1328125

1 0.1000000 0.1000000 0.1222222 0.1444444 0.2555556 0.2777778

BMI

Y [14.9,22.1) [22.1,24.1) [24.1,26.4) [26.4,29) [29,32.9) [32.9,49.1]

0 0.1601562 0.1679688 0.1718750 0.1757812 0.1640625 0.1601562

1 0.1555556 0.2000000 0.1111111 0.2222222 0.1666667 0.1444444

PREMENO

Y 0 1

0 0.8242188 0.1757812

1 0.7777778 0.2222222

MOMFRAC

Y 0 1

0 0.8867188 0.1132812

1 0.8111111 0.1888889

ARMASSIST

Y 0 1

0 0.6601562 0.3398438

1 0.5000000 0.5000000

SMOKE

Y 0 1

0 0.92187500 0.07812500

1 0.94444444 0.05555556

RATERISK

Y 1 2 3

0 0.3671875 0.3710938 0.2617188

1 0.1777778 0.4111111 0.4111111

FRACSCORE

Y [0,1) [1,3) [3,4) [4,6) [6,11]

0 0.10546875 0.32812500 0.12890625 0.26953125 0.16796875

1 0.05555556 0.13333333 0.10000000 0.30000000 0.41111111

**Appendix 3**

**Model 3 Output**

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

0 1

0.7589041 0.2410959

Conditional probabilities:

PRIORFRAC

Y 0 1

0 0.8050542 0.1949458

1 0.6136364 0.3863636

AGE

Y [55,58.6) [58.6,63) [63,68.8) [68.8,76.8) [76.8,90]

0 0.1696751 0.2093863 0.2057762 0.2057762 0.2093863

1 0.1136364 0.1136364 0.1704545 0.2045455 0.3977273

BMI

Y [14.9,22.4) [22.4,26.3) [26.3,30.8) [30.8,37.2) [37.2,49.1]

0 0.16606498 0.32851986 0.27075812 0.16245487 0.07220217

1 0.21590909 0.25000000 0.26136364 0.19318182 0.07954545

PREMENO

Y 0 1

0 0.8086643 0.1913357

1 0.8409091 0.1590909

MOMFRAC

Y 0 1

0 0.8880866 0.1119134

1 0.8295455 0.1704545

ARMASSIST

Y 0 1

0 0.6606498 0.3393502

1 0.4886364 0.5113636

SMOKE

Y 0 1

0 0.91335740 0.08664260

1 0.92045455 0.07954545

RATERISK

Y 1 2 3

0 0.3610108 0.3646209 0.2743682

1 0.1590909 0.4659091 0.3750000

**Appendix 4**

**Code output of R.**

|  |
| --- |
| > # Assignment 2  > # by Kenneth Lulie, Data 630 - Ami Gates  > # Created 10/12/2018  > # worked on 10/13/2018  >  > ##### Data preprocessing  >  > ##Load libraries in, as specified in the Naive Bayes Script provided  > #load the arules and e1071 library into memory  > #need to do this each time you start the new R session  > library("arules")  > library("e1071")  >  > #Read the data into a data frame  > origglow <- read.csv("D:/UMUC/630/Week 5/Assignment 2/glow500.csv")  >  > glow<-origglow  >  > #review data  > str(glow)  'data.frame': 500 obs. of 15 variables:  $ SUB\_ID : int 1 2 3 4 5 6 7 8 9 10 ...  $ SITE\_ID : int 1 4 6 6 1 5 5 1 1 4 ...  $ PHY\_ID : int 14 284 305 309 37 299 302 36 8 282 ...  $ PRIORFRAC: int 0 0 1 0 0 1 0 1 1 0 ...  $ AGE : int 62 65 88 82 61 67 84 82 86 58 ...  $ WEIGHT : num 70.3 87.1 50.8 62.1 68 68 50.8 40.8 62.6 63.5 ...  $ HEIGHT : int 158 160 157 160 152 161 150 153 156 166 ...  $ BMI : num 28.2 34 20.6 24.3 29.4 ...  $ PREMENO : int 0 0 0 0 0 0 0 0 0 0 ...  $ MOMFRAC : int 0 0 1 0 0 0 0 0 0 0 ...  $ ARMASSIST: int 0 0 1 0 0 0 0 0 0 0 ...  $ SMOKE : int 0 0 0 0 0 1 0 0 0 0 ...  $ RATERISK : int 2 2 1 1 2 2 1 2 2 1 ...  $ FRACSCORE: int 1 2 11 5 1 4 6 7 7 0 ...  $ FRACTURE : int 0 0 0 0 0 0 0 0 0 0 ...  > summary(glow)  SUB\_ID SITE\_ID PHY\_ID PRIORFRAC AGE WEIGHT HEIGHT BMI  Min. : 1.0 Min. :1.000 Min. : 1.00 Min. :0.000 Min. :55.00 Min. : 39.90 Min. :134.0 Min. :14.88  1st Qu.:125.8 1st Qu.:2.000 1st Qu.: 57.75 1st Qu.:0.000 1st Qu.:61.00 1st Qu.: 59.90 1st Qu.:157.0 1st Qu.:23.27  Median :250.5 Median :3.000 Median :182.50 Median :0.000 Median :67.00 Median : 68.00 Median :161.5 Median :26.42  Mean :250.5 Mean :3.436 Mean :178.55 Mean :0.252 Mean :68.56 Mean : 71.82 Mean :161.4 Mean :27.55  3rd Qu.:375.2 3rd Qu.:5.000 3rd Qu.:298.00 3rd Qu.:1.000 3rd Qu.:76.00 3rd Qu.: 81.30 3rd Qu.:165.0 3rd Qu.:30.79  Max. :500.0 Max. :6.000 Max. :325.00 Max. :1.000 Max. :90.00 Max. :127.00 Max. :199.0 Max. :49.08  PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE  Min. :0.000 Min. :0.00 Min. :0.000 Min. :0.00 Min. :1.00 Min. : 0.000 Min. :0.00  1st Qu.:0.000 1st Qu.:0.00 1st Qu.:0.000 1st Qu.:0.00 1st Qu.:1.00 1st Qu.: 2.000 1st Qu.:0.00  Median :0.000 Median :0.00 Median :0.000 Median :0.00 Median :2.00 Median : 3.000 Median :0.00  Mean :0.194 Mean :0.13 Mean :0.376 Mean :0.07 Mean :1.96 Mean : 3.698 Mean :0.25  3rd Qu.:0.000 3rd Qu.:0.00 3rd Qu.:1.000 3rd Qu.:0.00 3rd Qu.:3.00 3rd Qu.: 5.000 3rd Qu.:0.25  Max. :1.000 Max. :1.00 Max. :1.000 Max. :1.00 Max. :3.00 Max. :11.000 Max. :1.00  >  > #data seems good, no problems detected  >  >  > #this function shows results verifying the info given.  > apply(glow, 2, function (glow) sum(is.na(glow)))  SUB\_ID SITE\_ID PHY\_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE  0 0 0 0 0 0 0 0 0 0 0 0 0 0  FRACTURE  0  > #no NAS are detected  >  >  >  > summary(glow)  SUB\_ID SITE\_ID PHY\_ID PRIORFRAC AGE WEIGHT HEIGHT BMI  Min. : 1.0 Min. :1.000 Min. : 1.00 Min. :0.000 Min. :55.00 Min. : 39.90 Min. :134.0 Min. :14.88  1st Qu.:125.8 1st Qu.:2.000 1st Qu.: 57.75 1st Qu.:0.000 1st Qu.:61.00 1st Qu.: 59.90 1st Qu.:157.0 1st Qu.:23.27  Median :250.5 Median :3.000 Median :182.50 Median :0.000 Median :67.00 Median : 68.00 Median :161.5 Median :26.42  Mean :250.5 Mean :3.436 Mean :178.55 Mean :0.252 Mean :68.56 Mean : 71.82 Mean :161.4 Mean :27.55  3rd Qu.:375.2 3rd Qu.:5.000 3rd Qu.:298.00 3rd Qu.:1.000 3rd Qu.:76.00 3rd Qu.: 81.30 3rd Qu.:165.0 3rd Qu.:30.79  Max. :500.0 Max. :6.000 Max. :325.00 Max. :1.000 Max. :90.00 Max. :127.00 Max. :199.0 Max. :49.08  PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE  Min. :0.000 Min. :0.00 Min. :0.000 Min. :0.00 Min. :1.00 Min. : 0.000 Min. :0.00  1st Qu.:0.000 1st Qu.:0.00 1st Qu.:0.000 1st Qu.:0.00 1st Qu.:1.00 1st Qu.: 2.000 1st Qu.:0.00  Median :0.000 Median :0.00 Median :0.000 Median :0.00 Median :2.00 Median : 3.000 Median :0.00  Mean :0.194 Mean :0.13 Mean :0.376 Mean :0.07 Mean :1.96 Mean : 3.698 Mean :0.25  3rd Qu.:0.000 3rd Qu.:0.00 3rd Qu.:1.000 3rd Qu.:0.00 3rd Qu.:3.00 3rd Qu.: 5.000 3rd Qu.:0.25  Max. :1.000 Max. :1.00 Max. :1.000 Max. :1.00 Max. :3.00 Max. :11.000 Max. :1.00  >  >  >  > ### Removal of variables  > #remove the unique identifier for specific people  > glow$SUB\_ID<-NULL  >  > #remove the unique identifier for doctor, simply too many  > glow$PHY\_ID<-NULL  >  >  >  > #as we have no information about the six sites provided, we will also remove this.  > glow$SITE\_ID<-NULL  >  > #Fracscore is calculated off of age, priorfrac, momfrac, weight, armassist, and smoke  > #these are already included in our model (weight in is with BMI)  > #we need to remove, naive bayes works best with no correlation between variables  > glow$FRACSCORE<-NULL  >  > #remove weight and height, naive bayes assumes lack of correlation between independent variables  > #additionally, the relationship between this data is in the BMI variable  > glow$HEIGHT<-NULL  > glow$WEIGHT<-NULL  >  >  >  >  > ### CONVERT TO FACTORS  > glow$PRIORFRAC<- as.factor(glow$PRIORFRAC)  > glow$PREMENO<- as.factor(glow$PREMENO)  > glow$MOMFRAC<- as.factor(glow$MOMFRAC)  > glow$ARMASSIST<- as.factor(glow$ARMASSIST)  > glow$SMOKE<- as.factor(glow$SMOKE)  > glow$RATERISK<- as.factor(glow$RATERISK)  > glow$FRACTURE<- as.factor(glow$FRACTURE)  >  >  > ### Boxplot  >  > boxplot(glow$AGE, data = glow)  > boxplot(glow$BMI, data = glow)  > boxplot(glow$AGE, glow$BMI, col="maroon",names=c("Age at Enrollment","BMI at Enrollment"))  >  >  >  >  > ###Discretization  > #We will need to discretize age and BMI  > #Naive bayes requires discretization of continous variables  >  > #discretize age  > glow$AGE<-discretize(glow$AGE, "frequency", categories=6)  Warning message:  In discretize(glow$AGE, "frequency", categories = 6) :  Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!  > summary(glow$AGE)  [55,59) [59,63) [63,67) [67,72) [72,78) [78,90]  70 94 70 92 84 90  >  > #discretize the BMI variable  > glow$BMI<-discretize(glow$BMI, "frequency", categories=6)  Warning message:  In discretize(glow$BMI, "frequency", categories = 6) :  Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!  > summary(glow$BMI)  [14.9,22.1) [22.1,24.1) [24.1,26.4) [26.4,29) [29,32.9) [32.9,49.1]  84 83 83 83 82 85  >  >  >  >  >  >  >  >  >  >  > ##### Model building  >  > #make sure that the result is reproducible  > set.seed(1234)  >  >  > #Model 1  >  > #split the data into a training and test set  > ind <- sample(2, nrow(glow), replace = TRUE, prob = c(0.7, 0.3))  > train.data <- glow[ind == 1, ]  > test.data <- glow[ind == 2, ]  >  >  > #build the model and store in a variable model  > model<-naiveBayes(FRACTURE~., train.data)  > #output the model  > model  Naive Bayes Classifier for Discrete Predictors  Call:  naiveBayes.default(x = X, y = Y, laplace = laplace)  A-priori probabilities:  Y  0 1  0.7528409 0.2471591  Conditional probabilities:  PRIORFRAC  Y 0 1  0 0.7962264 0.2037736  1 0.5977011 0.4022989  AGE  Y [55,59) [59,63) [63,67) [67,72) [72,78) [78,90]  0 0.14716981 0.21132075 0.14339623 0.20377358 0.14339623 0.15094340  1 0.09195402 0.12643678 0.13793103 0.10344828 0.19540230 0.34482759  BMI  Y [14.9,22.1) [22.1,24.1) [24.1,26.4) [26.4,29) [29,32.9) [32.9,49.1]  0 0.1584906 0.1773585 0.1811321 0.1547170 0.1471698 0.1811321  1 0.1839080 0.1839080 0.1839080 0.1724138 0.1379310 0.1379310  PREMENO  Y 0 1  0 0.8000000 0.2000000  1 0.7701149 0.2298851  MOMFRAC  Y 0 1  0 0.8981132 0.1018868  1 0.8160920 0.1839080  ARMASSIST  Y 0 1  0 0.6490566 0.3509434  1 0.4597701 0.5402299  SMOKE  Y 0 1  0 0.92452830 0.07547170  1 0.93103448 0.06896552  RATERISK  Y 1 2 3  0 0.3471698 0.3811321 0.2716981  1 0.2298851 0.4137931 0.3563218  >  > #confusion matrix for the training set; need to round the estimated values  > table(predict(model, train.data), train.data$FRACTURE)    0 1  0 238 68  1 27 19  > #confusion matrix for the test data  > table(predict(model, test.data), test.data$FRACTURE)    0 1  0 101 29  1 9 9  >  > #mosaic plot  > mosaicplot(table(predict(model, test.data), test.data$FRACTURE), shade=TRUE, main="Predicted vs. Actual FRACTURES")  >  >  >  >  > ###Model 2  >  > ##Model 2 will reuse fracscore to see if it assists  >  > #Re-add Fracture score  > glow <- cbind(glow, FRACSCORE = origglow$FRACSCORE)  >  > glow$FRACSCORE<-discretize(glow$FRACSCORE, "frequency", categories=5)  Warning message:  In discretize(glow$FRACSCORE, "frequency", categories = 5) :  Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!  >  > ind <- sample(2, nrow(glow), replace = TRUE, prob = c(0.7, 0.3))  > train.data <- glow[ind == 1, ]  > test.data <- glow[ind == 2, ]  >  >  > #build the model and store in a variable model  > model<-naiveBayes(FRACTURE~., train.data)  > #output the model  > model  Naive Bayes Classifier for Discrete Predictors  Call:  naiveBayes.default(x = X, y = Y, laplace = laplace)  A-priori probabilities:  Y  0 1  0.7398844 0.2601156  Conditional probabilities:  PRIORFRAC  Y 0 1  0 0.8085938 0.1914062  1 0.5444444 0.4555556  AGE  Y [55,59) [59,63) [63,67) [67,72) [72,78) [78,90]  0 0.1484375 0.1992188 0.1406250 0.2148438 0.1640625 0.1328125  1 0.1000000 0.1000000 0.1222222 0.1444444 0.2555556 0.2777778  BMI  Y [14.9,22.1) [22.1,24.1) [24.1,26.4) [26.4,29) [29,32.9) [32.9,49.1]  0 0.1601562 0.1679688 0.1718750 0.1757812 0.1640625 0.1601562  1 0.1555556 0.2000000 0.1111111 0.2222222 0.1666667 0.1444444  PREMENO  Y 0 1  0 0.8242188 0.1757812  1 0.7777778 0.2222222  MOMFRAC  Y 0 1  0 0.8867188 0.1132812  1 0.8111111 0.1888889  ARMASSIST  Y 0 1  0 0.6601562 0.3398438  1 0.5000000 0.5000000  SMOKE  Y 0 1  0 0.92187500 0.07812500  1 0.94444444 0.05555556  RATERISK  Y 1 2 3  0 0.3671875 0.3710938 0.2617188  1 0.1777778 0.4111111 0.4111111  FRACSCORE  Y [0,1) [1,3) [3,4) [4,6) [6,11]  0 0.10546875 0.32812500 0.12890625 0.26953125 0.16796875  1 0.05555556 0.13333333 0.10000000 0.30000000 0.41111111  >  > #confusion matrix for the training set; need to round the estimated values  > table(predict(model, train.data), train.data$FRACTURE)    0 1  0 215 48  1 41 42  > #confusion matrix for the test data  > table(predict(model, test.data), test.data$FRACTURE)    0 1  0 98 24  1 21 11  >  > #mosaic plot  > mosaicplot(table(predict(model, test.data), test.data$FRACTURE), shade=TRUE, main="Predicted vs. Actual FRACTURES")  >  >  >  > ####Model 3  >  > #remove fracscore, rediscretize AGE and BMI with 8 ranges  > #remove fracscore  > glow$FRACSCORE<-NULL  >  > #Re add age and BMI  > glow$AGE <- origglow$AGE  > glow$BMI <- origglow$BMI  >  >  >  > #discretize age  > glow$AGE<-discretize(glow$AGE, "cluster", categories=5)  Warning message:  In discretize(glow$AGE, "cluster", categories = 5) :  Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!  > summary(glow$AGE)  [55,58.6) [58.6,63) [63,68.8) [68.8,76.8) [76.8,90]  70 94 103 118 115  >  > #discretize the BMI variable  > glow$BMI<-discretize(glow$BMI, "cluster", categories=5)  Warning message:  In discretize(glow$BMI, "cluster", categories = 5) :  Parameter categories is deprecated. Use breaks instead! Also, the default method is now frequency!  > summary(glow$BMI)  [14.9,22.4) [22.4,26.3) [26.3,30.8) [30.8,37.2) [37.2,49.1]  90 155 130 83 42  >  >  > ind <- sample(2, nrow(glow), replace = TRUE, prob = c(0.7, 0.3))  > train.data <- glow[ind == 1, ]  > test.data <- glow[ind == 2, ]  >  >  > #build the model and store in a variable model  > model<-naiveBayes(FRACTURE~., train.data)  > #output the model  > model  Naive Bayes Classifier for Discrete Predictors  Call:  naiveBayes.default(x = X, y = Y, laplace = laplace)  A-priori probabilities:  Y  0 1  0.7589041 0.2410959  Conditional probabilities:  PRIORFRAC  Y 0 1  0 0.8050542 0.1949458  1 0.6136364 0.3863636  AGE  Y [55,58.6) [58.6,63) [63,68.8) [68.8,76.8) [76.8,90]  0 0.1696751 0.2093863 0.2057762 0.2057762 0.2093863  1 0.1136364 0.1136364 0.1704545 0.2045455 0.3977273  BMI  Y [14.9,22.4) [22.4,26.3) [26.3,30.8) [30.8,37.2) [37.2,49.1]  0 0.16606498 0.32851986 0.27075812 0.16245487 0.07220217  1 0.21590909 0.25000000 0.26136364 0.19318182 0.07954545  PREMENO  Y 0 1  0 0.8086643 0.1913357  1 0.8409091 0.1590909  MOMFRAC  Y 0 1  0 0.8880866 0.1119134  1 0.8295455 0.1704545  ARMASSIST  Y 0 1  0 0.6606498 0.3393502  1 0.4886364 0.5113636  SMOKE  Y 0 1  0 0.91335740 0.08664260  1 0.92045455 0.07954545  RATERISK  Y 1 2 3  0 0.3610108 0.3646209 0.2743682  1 0.1590909 0.4659091 0.3750000  >  > #confusion matrix for the training set; need to round the estimated values  > table(predict(model, train.data), train.data$FRACTURE)    0 1  0 254 62  1 23 26  > #confusion matrix for the test data  > table(predict(model, test.data), test.data$FRACTURE)    0 1  0 93 29  1 5 8  >  > #mosaic plot  > mosaicplot(table(predict(model, test.data), test.data$FRACTURE), shade=TRUE, main="Predicted vs. Actual FRACTURES") |
|  |
| |  | | --- | | > | |