Naïve Bayes Analysis:

Burn Victims

David Hannon

UMGC Data630

Professor Ami Gates

**Introduction**:

**Objective**:

The purpose of this analysis is to create a naïve Bayes classification model that can accurately predict whether a burn victim entering a hospital will survive or die based on several other factors regarding their personal attributes as well as the nature of their burn injury.

**Problem Domain**:

Suffering from a severe burn injury is oftentimes a fatal occurrence. The American Burn Association estimates that the mortality rate is 69% for any patient who sustains an injury that covers over 70% of their total body surface area. This rate can also rise dependent on gender, age, and if injuries are also sustained through inhalation (burning of the throat, etc.) (NCBI, 2010). Those who survive severe burn injuries also can experience a difficult recovery process, commonly experiencing feelings of PTSD and depression.

Due to major advances in burn care made in the 1980s, survival rates have been on the rise ever since. According to Dr. David Herndon, “over the last 30 years at (his) burn center there has been a continuing reduction in the risk of mortality of about 2 percent per year in all age groups, burn sizes, and genders” referring to his burn units at the Shriners Hospitals for Children and University of Texas Medical Branch. He says specifically that the greatest advancement in survival rates has occurred in patients over the age of 40, stating that anyone over the age of 40 that sustains burns to 95% of their total body surface area will now survive half of the time (American College of Surgeons, 2018).

**Method Rationale**:

In order to evaluate the provided dataset and assess the probability of survival/death, the modelling technique of a naïve Bayes classification was chosen. A key contributor in deciding to use this modelling technique was that the variable of interest, Hospital Discharge Status, is a binary variable with either the outcome of “Alive” or “Dead”. In addition, many of the independent variables in this study are also binary, and the two non-binary variables can be easily discretized in order to have a more simplistic model output as well. While logistic regression may still be done on binary variables, the nature of the prediction being binary (0 or 1) lends itself more to the naïve Bayes output than the model equation of logistic regression.

**Analysis**:

**Data**:

The data used for this analysis was attained through GitHub and recommended through the Data630 course content. It describes 1000 different observations of patients that enter any of 40 hospitals to be treated for burns. The timeframe and specific locations of data collection are not made available. The variable of Hospital Discharge Status, the key dependent variable in this study, was marked as a 0 for “Alive” and 1 for “Dead”. Age was recorded as an integer for each patient. For Gender, 0 was recorded for “Female” and 1 is recoded for “Male”. For Race, 0 was recoded for “Non-White” and 1 was recorded for “White”. The Total Burn Surface Area, meaning the percentage of the body that sustained burns, was recorded as an integer describing the percentage from 0% to 100%. The variable of Inhalation Injury was recorded as a 1 if the patient sustained injury due to inhalation and as a 0 otherwise. Lastly the variable of Flame was recorded as a 1 if the injury involved a flame and a 0 otherwise.

**Exploratory Analysis**:

Using the “str” command, it was shown that R interpreted all variables as integers except for Age and Total Burn Surface Area, both of which were interpreted as numeric. This will be further addressed in pre-processing. Figure 1 shows the summary for the variable of the burn dataset (taken after pre-processing the integer variables). A picture containing holding, man, room

Description automatically generated

Figure 1

The information provided for the ID and Facility variables can be ignored as numerically speaking, these numbers are just identifiers. Keying in on Age as a numeric variable, it is important to notice that both the mean and median are in the lower 30s, despite the total range being from 0 to 90. This shows that this data is slightly skewed right, with most of the observations being in the left side for ages. Figure 2 confirms this shape as well, showing even further that the largest age group within our dataset is the 0-10 age group. Looking at the overall age range it is shown that the minimum age is .10, meaning infants are also included in this A screenshot of a cell phone

Description automatically generatedstudy. Prior knowledge states that younger age groups do have higher survival rates for burn injuries and thus in seeing this age distribution, it makes sense that Figure 1 also states that of our 1000 observed patients, 850 were discharged alive from the hospitals.

Figure

A screenshot of a cell phone

Description automatically generated Examining the other numeric variable of Total Burn Surface Area (TBSA), once again the mean and median are low (13.54% and 6.00% respectively) despite the overall range covering from .10% to 98% TBSA. Figure 3 shows the histogram that very much emphasizes the rightward skew of this variable. Once again, knowing that most of these 1000 observed burn patients did not experience what is considered severe burns (TBSA>40%) may be a major factor in why the data shows that only 15% of patients did not survive their burn injuries.

Figure

A screenshot of a cell phone

Description automatically generated With such strong skew for the TBSA variable, it was important to visualize if there were any outliers within the dataset for extreme percentages of TBSA. Figure 4 shows a boxplot that showcases these outliers. Because of how clustered the data is at the lower end of the number line, the upper fence for outliers is set just below 40% TBSA, meaning all patients who are deemed as “severe” burn patients are considered to be outliers within this particular dataset. After some pre-processing techniques were used, it was shown that 81 observations recorded a TBSA over 40%. Due to the quantity of outliers and the importance of including those deemed as “severe” burn patients, the decision was made to not remove outliers. A similar boxplot was created for the variable of Age but yielded no outliers.

Figure

Observing the binary data from the summary command in Figure 1, it can be seen that 70.5% of patients were male and 29.5% were female, 58.9% of patients were white and 41.1% were non-white, 87.8% of patients did not experience an inhalation injury, and 52.9% of the observed patients’ injuries involved a flame.

**Preprocessing**:

In order to pre-process the data, the first step was to remove any unnecessary variables. The first unnecessary variable was the ID variable as it simply gives a numeric ID to each observation which has no meaning other than for differentiation. The second variable deemed as unnecessary for this modelling technique was the Facility identifier. This data was collected from 40 different facilities, and while it may be interesting to do a different analysis to see which hospitals/facilities are getting better results in terms of survival rates, for the naïve Bayes classification, this variable would need to be discretized which does not make sense given that the numbers assigned to each hospital is arbitrary. Both of these variables were set to null in order to eliminate them from the dataset.

After unnecessary variables were eliminated, it was imperative to look for any bad observations. As previously discussed, outliers were found within the variable of TBSA but the choice was made to maintain these observations for accuracy and inclusiveness. Utilizing the output of the summary command, it was shown that the dataset had no blank observations as these would have appeared as an “NA” within Figure 1.

The next step in pre-processing was to address that R interpreted all of the binary variables in the dataset as integer variables rather than as factor variables. In order to successfully run the naïve Bayes analysis, all variables must be recognized as factors, thus the factor function was used on the variables of Death, Gender, Race, Inhalation Injury, and Flame. Once running this, the “str” command was run to check that these variables were corrected to factors. Lastly, the two numeric variables of Age and TBSA needed to also be converted to factor variables using the discretization function. It was decided to use the fixed method of discretization to ensure analyst control within the groupings to emphasize age groupings as well as cut offs for severity levels within TBSA. The age variable was discretized into 10-year age ranges to align with the histogram visualization already created in Figure 2. The TBSA variable was discretized by intervals of 20%, starting at 0%, to both ensure that the 40% level was a clear cutoff between groupings since this is the defining percentage for “severe burns”, but also to ensure that groupings would be sufficiently large in the higher range to include at least 24 observations within each. The “summary” command was used to ensure that all of the numeric data was fit into one of the groups and that no NA’s were created.

**Algorithm Intuition**:

In order to have a large enough sample from the data to create an accurate model as well as a separate and large enough sample on which to test the accuracy of the model, the dataset of 1000 patient observations was split roughly 70/30, yielding 698 observations within the training dataset and 302 observations within the testing dataset. Using the training dataset, the “naiveBayes” command was used, with its output stored as one of three models. This function requires the e1071 package through R and takes the input of a dataset, a dependent variable, and all independent variables. In this case, these inputs are the training dataset, the binomial factor of hospital discharge status (either 0 or 1 for “Alive” or “Dead”) labelled “DEATH”, and the different independent variables for each model. Model 1 utilized the “.” shortcut to include all independent variables, Model 2 input only the variables of Age, TBSA and Gender, while model 3 input only Gender, Race, and TBSA. The purpose of creating these three models was to then utilize each in creating a confusion matrix to compare accuracies.

**Modeling Fitting**:

In order to evaluate levels of accuracy amongst the three created models, Confusion matrices were completed for each using both the training dataset and the test dataset. Figure 5 shows the Confusion Matrices for all 3 models using the training data set. Model 1, based on all available independent variables, yielded 643 correct predictions from a total of 698 observations, showing 92.12% accuracy. Model 2, based on Age, TBSA, and Gender, yielded 652 correct predictions, showing 93.41% accuracy. Model 3, based on Race, Gender, and TBSA, yielded 637 correct predictions, showing 91.26% accuracy. While all seem very close in accuracy, Model 2 had the highest predictive strength when using the training data set. A similar analysis of accuracy on test data also supports Model 2 as having the strongest predictive strength. Because of this, Model 2 was the selected model to be applied for the results.

A picture containing bird

Description automatically generatedA screen shot of a person

Description automatically generatedA picture containing room, remote

Description automatically generated

Figure

**Result**:

**Output and properties**:

Figure 6 shows the output of Model 2, which shows conditional probabilities for Age, TBSA, and Gender broken down with whether the patient was discharged as alive or dead. The A-priori probabilities show that the model assumes 84.96% of patients survive and 15.04% do not, based on the 850 out of 1000 that survived within our dataset. The rows within the conditional probabilities represent either alive (0) or dead (1). For example, the 0 row listed under TBSA can be interpreted as such: Given that a patient survived, there is an 88.36% chance their TBSA was between 0 and 20 exclusive, a 10.11% chance their TBSA was between 20 inclusive and 40 exclusive, a .51% chance their TBSA was between 40 inclusive and 60 exclusive, a .67% chance their TBSA was between 60 inclusive and 80 exclusive, and a .34% chance their TBSA was between 80 inclusive and 100. This certainly aligns with the definition for severe TBSA as very few of the survivors were patients with TBSA above 40. While reading the 1 row, or death probabilities for TBSA, it is not as obvious of the severity of such high TBSA but this is because of the skewedness of the data containing so many more lower observations and increasing the quantity of patients who died from less severe TBSA.

While the naïve Bayes classifiers do not lend themselves to writing out one formal equation due to its reliance on categorical variables, a probability for survival can be calculated for any patient who enters the hospital by accounting for their age, TBSA, and gender. In order to do this one must multiply the conditional probability for each known factor with the overall conditional probabilities and then divide by the A-priori probability to then receive a probability of either death or survival depending on which was chosen to be calculated. A screenshot of a cell phone

Description automatically generated

Figure

**Evaluation and Diagnostics**:

A picture containing holding

Description automatically generatedFigure 7 shows a confusion matrix for Model 2, now based on the test data rather than the training data. It shows that the model correctly predicted that 252 of the actual survivors would survive while it predicted that 23 of the actual survivors would die. It also correctly predicted that 22 of those who died would die but predicted that 5 of those who died would actually survive. Overall this is a precision of 90.73% which is fairly reliable.

Figure

A screenshot of a cell phone

Description automatically generatedFigure 8 shows a mosaic plot of this same confusion matrix. Based on the shadings, one can see that the red shadings imply when the model predicted the incorrect outcome for a patient. However, the shading of light red implies that the residual was not extremely far off in its prediction as that would be shown by a darker red color. As for the large white rectangle, this implies that were no residuals for the models correctly predicted survivals and the dark blue implies a large residual for the models correctly predicted deaths.

Figure

Figure 9 shows the ROC curve for this particular model. Thinking of this in terms of a null hypothesis, one would expect that a model that randomly assigns the prediction of “Alive” or “Dead” in this case would be true 50% of the time and false 50% of the time. That is what the grey, dashed line represents. The calculated ROC curve shows that this model is closer to the top left corner which implies strong performance as the model is much more likely to yield a true-positive than a false-positive. A close up of a map

Description automatically generated

Figure

**Conclusion**:

**Summary**:

Using the naïve Bayesian classification, three models were created intended to be able to predict the probability of either life or death given a burn victim’s personal attributes and injury details upon their arrival at a treatment facility. The strongest output model only required to know a patient’s Age, Total Burn Surface Area, and Gender to be able to correctly predict whether that patient would survive or not an estimated 90.73% of the time. Having a model of this nature could very much help people such as EMT’s and Emergency Room staff in determining the urgency levels of patients who have been burn victims. Being able to prioritize those patients who are predicted most likely to die may further improve the overall burn survival rates which have already been on the rise.

**Limitations and Improvements**:

While some of the naïve Bayes classification’s outputs do give us insight as to which factors increase a patient’s chances at survival, the overall model does not specifically point these things out and thus the model may not be helpful in improving treatment methods based on certain characteristics. However, follow up models of liner or logistic regression could be done to pinpoint these factors. Unfortunately, the naïve Bayes method does not include anything comparable to p-values for each variable to be able to compare significance in this way.

Similarly, this specific type of analysis does not have a technique comparable to the “minimal adequate model” like a linear regression or logistic regression techniques may use. It is for this reason that the analyst chose to experiment with three different models to compare accuracy, however, with 6 different variable to either choose to include or not, there are realistically 60 other possible combinations of independent variables that could have been chosen to create a model and perhaps one of these models may yield higher accuracy than the model chosen for this analysis. Thus, for improvement, the analyst would either utilize a new modelling technique that automates R to run through all of the possible model combinations, or potentially take the time to determine which of all the possible combinations were actually the strongest.

**References:**

Garneski, S., & Rose, D. (n.d.). Burn Specialists Report a Dramatic Increase in Burn Injury Survival Over the Past 30 Years. Retrieved June 22, 2020, from https://www.facs.org/media/press-releases/2018/burninjury030918

Pavoni, V., Gianesello, L., Paparella, L., Buoninsegni, L., & Barboni, E. (2010, April 27). Outcome predictors and quality of life of severe burn patients admitted to intensive care unit. Retrieved June 22, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873368/