Assignment 2

Burn Victim Analysis

Data 630-9040 Machine Learning

Prepared by Kierra Young

Prepared for Professor Bati

6/21/2021

**Introduction**

**Objective**

The objective of this analysis is to predict what variables are more likely to a person dying from burns. The type of analysis this is exploratory. This analysis will analyze age, gender, if the person was burned by a flame, inhaled smoke, the facility they went to for treatment, or had a surface burn to their skin. The specific modeling that will be used will be a bar plot for visualization, logistic regression which will guide the plot classification, and a confusion matrix. There will be more methods of evaluation to verify the measure of precision when predicting the chance of a person dying from certain types of burns depending on the variables mentioned above.

**Problem Domain**

Some interesting facts on burn victims according to the NCBI is that

“Five-thousand two-hundred-sixty patients were admitted after burn injury from July 1989 to June 2009, and of those, 145 patients died after burn injury. Of these patients, 144 patients had an autopsy. The leading causes of death over 20 years were sepsis (47%), respiratory failure (29%), anoxic brain injury (16%), and shock (8%). From 1989 to 1999, sepsis accounted for 35% of deaths but increased to 54% from 1999 to 2009, with a significant increase in the proportion due to antibiotic resistant organisms (P < 0.05).”(ncbi.com, 2009)

This is just some statistical background information on burn victims who have passed away due to several reasons. The dataset being used to preform this analysis will be able to examine if the information above is statistically true or if there may be other factors that lead to a greater chance of a person dying from a certain type of burn.

**Method Rationale**

The method of rationale that I will be using is relevant to the finding the reason for burn victim deaths. Because with the exploratory analysis, I will be able to help a hospital or facility better understand what the problem maybe when it comes to certain burn victims dying versus the ones who survive. Likewise, using machine learning techniques such as logistic regression will be used to find the accuracy of the models that will be analyzed.

**Analysis**

**Data**:

There are 1000 obs. of 9 variables. The person who owns this dataset name is lbraglia /aplore3 on github. He or she is based in Italy. The names of the variables in this dataset are ID, FACILITY, DEATH, AGE, GENDER , RACEC, TBSA, INH\_INJ, and FLAME. TBSA stands for total burn surface area and INH\_INJ stands for burn involved inhalation injury. Some of the variables in the dataset are numerical and others are labeled integers.

**Exploratory Analysis**:

**Str**

The str function gives an overview of the burn data framework. The output describes the type of variables in the data frame. All the variables were labeled as int and num. There were 1000 observations for 9 different variables. In addition, it also shows the number of variables and data in the data frame individually. The target variable that will be analyzed is “death” and the other variables mentioned in the data section are the independent variables that will help indicate a burn victims’ chance at death.

**Summary**

The summary portion gives an overview of the statistics for each variable in the burn dataset. The output for the burn dataset was easily placed into min, max, quarters, and medians for all the variables. There also were not any NA values listed on the summary for the variables so all the data is there so no removal will be necessary.

Bar Plot

Figure 1. in the Appendix is a bar plot of the variables flame versus inh\_inj. This diagram shows the status of the amount of people in the dataset who were burned more severely with either flame or because of inhalation burns. From the diagram people with flame or not flame was highly likely to visit a facility and people who had inhalation burns only came in because they inhaled the fire smoke and was also burned internally. This bar plot is beneficial for when later observing deaths because it gives the hypothesis that more people come into a facility from flame burns but also majority of people who visit a facility who were listed under inh\_inj were highly likely to have been burned.

**Preprocessing**

From observing str and summary all the values were in the dataset mostly were int and num. Because of this no discretion was needed, however there was a need to prepare the commands for the logistic regression modeling. Likewise, to prepare the data for logistic regression. The regression was used to split the data into a training and test set. This was so that 70% of the data was used for training and the other 30% was used for precision narrowing then judged for accuracy. After that was preprocessed, the logistic regression was soon able to be performed.

**Algorithm Intuition**

As for algorithm logic, I decided to use a logistic regression model. This model will be able to predict a certain percentage of how likely each of the variables are to a burn victim dying once reaching a facility. The key parameters of the algorithm will display the coefficient, p value, and standard error for each independent variable that possibly leads to a victim’s death. The equation for the logistic regression algorithm is model<-glm(DEATH~., family=binomial, data=train.data). Glm stands for general linear model.

**Model Fitting**:

The key steps and activities used to fit the model were to determine the target variable in the dataset that I wanted to examine. The target variable I chose was death. To tune the model, I used a code that could split the data into training and test data. “We use the training set to build the classification tree model, and we use the test set to evaluate the accuracy of a model” (Logistic Regression). The training data was pruned to only 70% while the test was pruned to 30%. This was done to see if the sample amount would or could preform highly when applying the logistic regression. After doing this for the glm the input parameters were formula, family, and dataset name. These parameters were put into the equation above in the algorithm section to be stored under the name “model” for training the data.

For fitting the model, the “fitted.values model stores the predicted probabilities”(Logistic Regression) of positive burn experimental findings for each case in the training set. To fit the model, I used a command that shows the predicted values for the first 10 values on the training set.

**Result**

Algorithm glm Results

The results of the algorithm glm’s output is shown as figure 4 in the appendix. The results show that the coefficient for facility, inh\_inj, and flame variables are negative. If we increase those three variables by 1 and maintained the values of the leftover independent variables continuous, according to Logistic regression docs the odds of positive death for burns would decrease by 0.5022 for facility, 1.8852 for inh\_inj, and 9.2309 for flame.

However, if we increase the variables age, gender, race, and tbsa by 1 and leave the values of the remaining independent variables endless, the odds of death for burns would increase by 0.2263 for age, 9.4487 for gender, 8.5327 for race, 0.4671 for tbsa. To interpret, the algorithms output means that with age, gender, race, and tbsa if those circumstances are raised one is more likely that a patient died from burns while facility, inh\_inj, and flame increasing by 1 a patient is less likely to die from burns in this dataset. The stated objective has been met because it shows what variables are more likely to cause someone to die when being brought into a burn facility.

**Model Properties**

The characteristic or goal of the fitted model is that it displays the overall predicted probabilities for the training data section. The command for the fitted model displays the patient id and the subsequent projected likelihood for the first 10 occurrences in the burn training set. In addition, the command “fitted.values” for the burn dataset holds the forecast odds of positive deaths from train.data test results for each case in the training set.

**Evaluation**:

The confusion matrix helps to evaluate the performance of the fitted model. Both the train and test data were evaluated by using the matrix. In addition, the confusion matrix output is located as figure 3 in the appendix. To begin, the test data output was able to give the accuracy of the model. It gave the results that the number of correctly classified instances were 301. Then the number of wrong instances was 1. The equation was 301/302 which equaled 99.62% classification accuracy for the testing section. In addition, for the training data section the number of correctly classified instances were 698/699 which equaled 98.57% accuracy for the training section.

**Diagnostics**:

From observing the diagnostics plot. There is an extremely high prediction that with the model formed that the variables (age, gender, race, tbsa)mentioned above will be the main indictors of a person dying from a burn. In contrast the variables (facility, flame, inh\_inj) will also be a main indictor for person to more than likely not die burns. The residual plot shows the blue lines in a vertical fashion. One it is pointing all the way up for positive chance of death and the others are pointing directing down indicating a negative chance of death. The black line is strictly in the middle, so the prediction value is extremely high.

Furthermore, Figure 5 in the appendix shows an AIC model for prediction of the main causes related to burn deaths. The strongest variable indicator was TBSA. Which means total burn surface area is the strongest prediction for when a person will die from a burn. In relation to the logistic regression model, the variables DEATH ~ ID + AGE + GENDER + RACEC + TBSA ranked as the 12 AIC to be the main indictors of death. The variables flame, facility, and inh\_inj were removed before them which means the logistic regression and confusion matrix accuracy was extraordinarily strong.

**Conclusion**

**Summary**:

An especially important highlight for this analysis is that flame, facility, and inh\_inj have exceedingly small correlation to a person dying from a burn. Even though it may seem as though inhalation of smoke that can cause burns, a flame from a fire, and a choosing a good facility may be the most understandable reason for a person dying from a burn; those have been found not to be the leading causes of burn death. A burn caretaker, burn facility, or even a person who does not know how to take care of a burn victim can benefit from this analysis, because we know what really indicates if a person will die from a burn. It was found that the strongest indicator of a person dying from a burn is total burn surface area. The other principles that follow behind are age, gender, and race. The total accuracy of this analysis has a p- value of 0.95%. The confusion matrix had an accuracy of 98% for the training data and 99 % for the test data. One interesting finding from an article called “Burn Data Shows Who is Most Affected by Burns” to take away is that it says “Mortality rate increases as TBSA increases. Victims with burns that encompass 70 percent or more of their TBSA have a 55 percent chance of dying from their wounds, and the percentage increases as TBSA does. Overall, the burn mortality rate was 3.3 percent for all cases.”(erieinsurance.com)

**Limitations**:

Some limitations of the analysis were making a histogram and scatter plot. Some of the codes were having difficulty adjusting to the framework of the dataset. If may have just been that the codes needed to be adjusted more or the dataset was not fitting well with some of those given codes. There were not many issues with the algorithm except for there was a warning message that the glm did not converge, however it did show the model summary so that was slightly confusing. One other limitation was the plot residual and the confusion matrix. The accuracy was extremely high so I am not sure if the data may have been leveraged in that way by the people who put collected but accuracy was a little high. An interesting phenomenon to discuss more. However, those were my main limitations with this dataset.

**Improvement Areas**

Some potential areas for my improvements would be to find more options in code when making histograms and box plots for observing the dataset. To learn how to manipulate the code better in the histogram and box plots so it can adjust with the dataset.

References

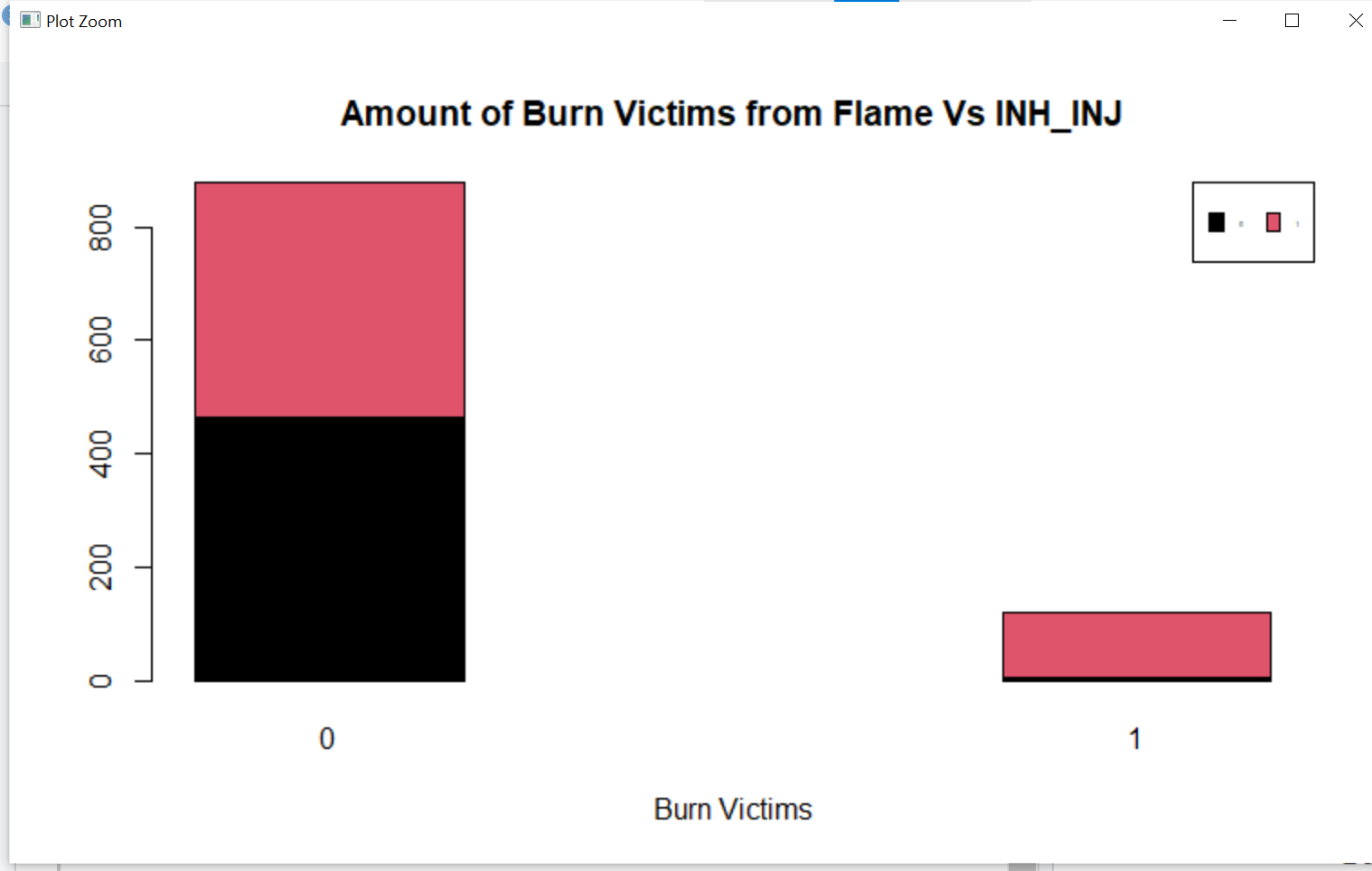
Insurance, E. (2015, January 29). Burn Data Shows Who is Most Affected by Burns. erieinsurance.com. https://www.erieinsurance.com/blog/burn-data-shows-affected-burns.

UMGC. (n.d.). Logistic Regression R. Aldephi.

Williams, F. N., Herndon, D. N., Hawkins, H. K., Lee, J. O., Cox, R. A., Kulp, G. A., Finnerty, C. C., Chinkes, D. L., &amp; Jeschke, M. G. (2009). The leading causes of death after burn injury in a single pediatric burn center. Critical care (London, England). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811947/.

Appendix

**Figure 1: Bar Plot- 0=no, 1= yes**

******

**Figure 2: Algorithm glm output:**

#build the model and store in a variable model

> model<-glm(DEATH~., family=binomial, data=train.data)

Warning messages:

1: glm.fit: algorithm did not converge

2: glm.fit: fitted probabilities numerically 0 or 1 occurred

> #output the coeficients and Residual Deviance

> print(model)

Call: glm(formula = DEATH ~ ., family = binomial, data = train.data)

Coefficients:

(Intercept) ID FACILITY AGE

-1629.5479 1.8987 -0.5022 0.2263

GENDER RACEC TBSA INH\_INJ

9.4487 8.5327 0.4671 -1.8852

FLAME

-9.2309

Degrees of Freedom: 697 Total (i.e. Null); 689 Residual

Null Deviance: 591.1

Residual Deviance: 1.711e-06 AIC: 18

**Figure 3:Confusion Matrix Code and Output**

#confusion matrix for the training set; need to round the estimated values

> table(round(model$fitted.values), train.data$DEATH)

0 1

0 593 0

1 0 105

> table(round(predict(model, train.data, type="response")), train.data$DEATH)

0 1

0 593 0

1 0 105

> #display the first 10 estimated values for the test data

> predict (model, test.data, type="response")[1:10]

5 14 16 26 28

2.220446e-16 2.220446e-16 2.220446e-16 2.220446e-16 2.220446e-16

29 36 39 40 50

2.220446e-16 2.220446e-16 2.220446e-16 2.220446e-16 2.220446e-16

> #store the estimated values in a variable mypredictions; need to round the values

> mypredictions<-round(predict (model, test.data, type="response"))

> #confusion matrix for the test data

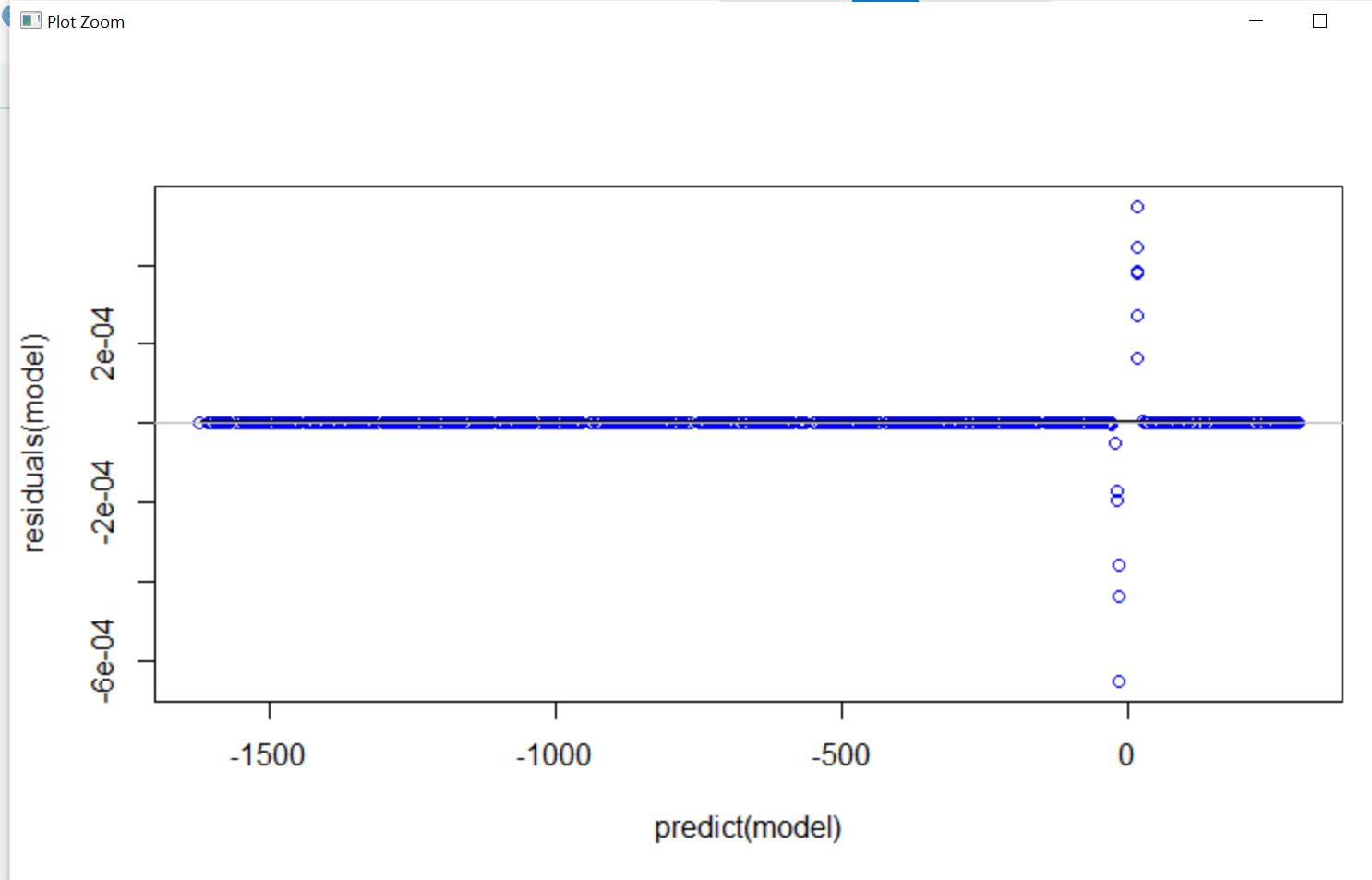
> table (mypredictions, test.data$DEATH

mypredictions 0 1

0 256 0

1. 1 45

**Figure 4: Residual Plot**

****

**Figure 5: AIC Output**

> #minimal adequate model

> summary(step(model))

Start: AIC=18

DEATH ~ ID + FACILITY + AGE + GENDER + RACEC + TBSA + INH\_INJ +

FLAME

Df Deviance AIC

- INH\_INJ 1 0.00 16.00

- FLAME 1 0.00 16.00

- RACEC 1 0.00 16.00

- GENDER 1 0.00 16.00

- FACILITY 1 0.00 16.00

- AGE 1 0.00 16.00

- TBSA 1 0.00 16.00

<none> 0.00 18.00

- ID 1 218.39 234.39

Step: AIC=16

DEATH ~ ID + FACILITY + AGE + GENDER + RACEC + TBSA + FLAME

Df Deviance AIC

- FLAME 1 0.00 14.00

- FACILITY 1 0.00 14.00

- RACEC 1 0.00 14.00

- GENDER 1 0.00 14.00

- AGE 1 0.00 14.00

- TBSA 1 0.00 14.00

<none> 0.00 16.00

- ID 1 229.63 243.63

Step: AIC=14

DEATH ~ ID + FACILITY + AGE + GENDER + RACEC + TBSA

Df Deviance AIC

- FACILITY 1 0.00 12.00

- RACEC 1 0.00 12.00

- GENDER 1 0.00 12.00

- AGE 1 0.00 12.00

- TBSA 1 0.00 12.00

<none> 0.00 14.00

- ID 1 239.16 251.16

Step: AIC=12

DEATH ~ ID + AGE + GENDER + RACEC + TBSA

Df Deviance AIC

- GENDER 1 0.00 10.00

- AGE 1 0.00 10.00

- RACEC 1 0.00 10.00

- TBSA 1 0.00 10.00

<none> 0.00 12.00

- ID 1 239.22 249.22

Step: AIC=10

DEATH ~ ID + AGE + RACEC + TBSA

Df Deviance AIC

- RACEC 1 0.00 8.00

- AGE 1 0.00 8.00

- TBSA 1 0.00 8.00

<none> 0.00 10.00

- ID 1 240.32 248.32

Step: AIC=8

DEATH ~ ID + AGE + TBSA

Df Deviance AIC

- AGE 1 0.00 6.00

- TBSA 1 0.00 6.00

<none> 0.00 8.00

- ID 1 241.38 247.38

Step: AIC=6

DEATH ~ ID + TBSA

Df Deviance AIC

- TBSA 1 0.00 4.00

<none> 0.00 6.00

- ID 1 369.86 373.86

Step: AIC=4

DEATH ~ ID

Df Deviance AIC

<none> 0.00 4.00

- ID 1 591.14 593.14

Call:

glm(formula = DEATH ~ ID, family = binomial, data = train.data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.488e-03 -2.000e-08 -2.000e-08 -2.000e-08 2.433e-03

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5394.508 104651.516 -0.052 0.959

ID 6.354 123.271 0.052 0.959

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5.9114e+02 on 697 degrees of freedom

Residual deviance: 1.2111e-05 on 696 degrees of freedom

AIC: 4

Number of Fisher Scoring iterations: 25