Logo

Description automatically generated

Week 3 Practice

Craig Perkins – [perkins.cr@northeastern.edu](mailto:perkins.cr@northeastern.edu)

Northeastern University

November 2022 – ALY 6150

**Week 3 Practice  
  
Introduction**

In this assignment I analyze data about hospital cases to build models to predict whether a case is likely to be severe or not using logistic regression and build a separate regression model to predict length of stay. The dataset has 18 features and 318,438 records. The dataset includes demographic information on the patient, high level information about the hospital and also contains information about the visit like the type of admission. This analysis investigates confounding variables for epidemiological analysis and the importance of features for analysis. I found this exercise to be challenging to produce a good model and plan to revisit this again in the future to produce a more in depth analysis.

1. **Provided descriptive statistics tables for important variables.**

Taking a preliminary look at the data, it appears that all columns are important variables in this data. I removed the column named `Unnamed: 0` which comes in from not removing the index when creating a CSV file, but left all other columns in. I may revisit this later after more analysis, but each column looks important except for `case\_id`.

Graphical user interface, application

Description automatically generated

**2. Compute the OR of Age on Severity.of.Illness = "Extreme" or not as the outcome. Note that Age is a categorical variable. You should use Age=21-30 as the base. Alternatively, you may convert Age categories into one numeric age variable by taking the mid-point of the age ranges, i.e. ages 11-20 might be recoded to 16.**

|  |  |  |
| --- | --- | --- |
| Severity of Illness / Age | 21 - 30 | Other |
| Extreme | 1358 | 55365 |
| Not Extreme | 15410 | 246305 |

Odds Ratio = (a / c) / (b / d) = (1358 / 15410) / (55365 / 246305) = 0.39

1. **Repeat (compute the OR of Age) while controlling for 5 or 10 confounding variables (i.e. additional controls in a logistic regression). What happened to the OR of Age categories?**

For this analysis I chose to control:

* (df['Type of Admission'] == 'Trauma')
* (df['Department'] == 'gynecology')
* (df['Hospital\_region\_code'] == 'X')
* (df['Ward\_Type'] == 'R')
* (df['Hospital\_type\_code'] == 'a')

Which resulted in 17014 rows for analysis controlling for the confounding variables.

|  |  |  |
| --- | --- | --- |
| Severity of Illness / Age | 21 - 30 | Other |
| Extreme | 191 | 1998 |
| Not Extreme | 2481 | 12344 |

OR after adjusting for confounding variables:

Odds Ratio = (a / c) / (b / d) = (191 / 2481) / (1998 / 12344) = 0.476

I have seen many pages only about a 10% rule of thumb for confounding variables to determine whether a variable is important enough to include in developing a model.

1. **Use statistical learning or machine learning to predict the Severity.of.Illness = "Extreme" or not. Provide appropriate visualization, confusion matrix, and other metrics as necessary.**

In order to build a logistic regression model, I had to perform a significant amount of pre-processing and feature engineering to prepare the data.set. The pre-processing I performed includes

* Imputing missing data with the median
* Calculating an average of the Age and Length of Stay ranges to make the values numeric. For this I took the sum of the low end of the range and high end of the range. For length of stay there were values “More than 100 days” and for this I replaced the cells with 100-999 so I could apply the (low + high) / 2 logic.
* One hot encoding on categorical data
* Scaling of numerical columns to keep the values in each series of the dataframe between 0 and 1.

After all of the pre-processing I performed a train-test split (25% train and 75% test) and created an untuned logistic regression to see how accurate the model would be without any hyper parameter tuning. I ended up getting a model that was 82% accurate and produced the following confusion matrix:

Graphical user interface, text, application, email

Description automatically generated

Below is the ROC curve of the model:

Graphical user interface, text

Description automatically generated

After completing the rest of the exercises in the assignment, I plan to come back to the logistic regression to see if I can improve upon the model because I did not find these results satisfying. The model should be predicting true more often.

1. **Do similar exercise using length of stay (LOS). Please note LOS will need to be converted into a single numerical variable to make predictive analytics plausible. How do you visualize this?**

For this problem I chose to use a DecisionTreeRegressor for its explainability and ease of visualization. For this model I chose a max\_depth of 4 to keep the tree shallow and to see what the most important features that the model selects are.

**Graphical user interface, text, application, email

Description automatically generated**

Interestingly for the DecisionTreeRegressor I found that the number of visitors for the patient was the most important feature when determine length of stay. I found the result surprising initially, but after a while it made sense to me. The longer a patient is in the hospital, the more people tend to go visit the patient to wish them better. I think it’s interesting how tree-based models select the most important features and provide a way to visualize how a decision is ultimately made. See below for an image of the tree, for the top 2 levels the branching left and right is done on the amount of visitors to the patient.

**A picture containing text, boat

Description automatically generated**

**Conclusion**

This analysis was a good exercise to practice analysis for healthcare using a rich example dataset. The dataset contains many problems that data scientists will face with real datasets like preparing the data for analysis, engineering features and iterating through different models to find a model that makes the best trade-offs for the problem at hand. The logistic regression model was not accurate in this case and I am wondering if I made a mistake when processing the dataset and should perform the analysis again with different permutations of the original dataset and features to include in the final model. The regression model with decision trees really helped with explainability to explore how the decision tree chose to make a decision. In my personal experience working in healthcare, many doctors like the explainability of a decision tree model yet there are shortcomings of decision trees that make other methods like random forest and gradient boosting preferable to decision trees.