Lab 1 - PECARN TBI Data, STAT 214, Spring 2025

February 23, 2025

1 Introduction

Traumatic brain injury (TBI) is one of a major cause of death and disability in children worldwide. Identifying children at low risk for clinically important TBI (ciTBI) is significant to avoid unnecessary CT scans, which may cause additional damage to children. Therefore, training a model to identify patients who are necessary to the CT scan is meaningful. With the TBI data, we can discovery the potential relationship between clinical observations and ciTBI based on the exploratory data analysis and model coefficients. In this report, the arrangement of the content is as follows:

First, we will introduce the data source in Section2.1. Then, we will focus on the data exploration and cleaning in Section2.2 and Section2.3, as well as the findings during the process in Section3. After cleaning the data set, we will fit two predictive models in section4. Finally, we provide our discussion and conclusion.

2 Data

In this section, we will focus on the data exploration and data cleaning, which are also the most important part of this project.

2.1 Data Collection

The dataset we used in this project is provided by Kuppermann et al. In their work, they enrolled patients younger than 18 years presenting within 24 h of head trauma with Glasgow Coma Scale scores of 14–15 in 25 North American emergency departments, aiming at deriving an age-specific prediction rules for ciTBI (death from traumatic brain injury, neurosurgery, intubation $>$24 h, or hospital admission $\geq$2 nights), which can be used to identify children at very low risk of ciTBI for whom CT might be unnecessary and avoid the potential negative influence. There are 42,412 samples and 125 variables in total, among which 763 patients did have ciTBI. The variables collected are mainly the mechanism of injury, clinical variables (history and symptoms), and physical examination findings. With these variables, we are going to predict whether a patient is under high ciTBI risk and if it is necessary to take the CT test.

2.2 Data Exploration

- The main goal of this section is to give the reader a feel for what the data “looks like’ ’ at a basic level
- Think about plots that summarize the data, plots that convey some smaller findings which ultimately motivate your main findings
- A good report will tie everything together so that there is a reason for every figure in the story

In this section, we will visualize the distribution of some key variables to show the data set in a basic level and give a glance at it. We will not cover the findings or data cleaning. For more details please see Section 2.3 and Section 3.

Most variables in the data set are categorical, and can be divided into several groups. For instance, AMS indicates whether GCS $<$ 15 or other signs of altered mental status, and AMSAgitated, AMSSleep, AMSSlow,

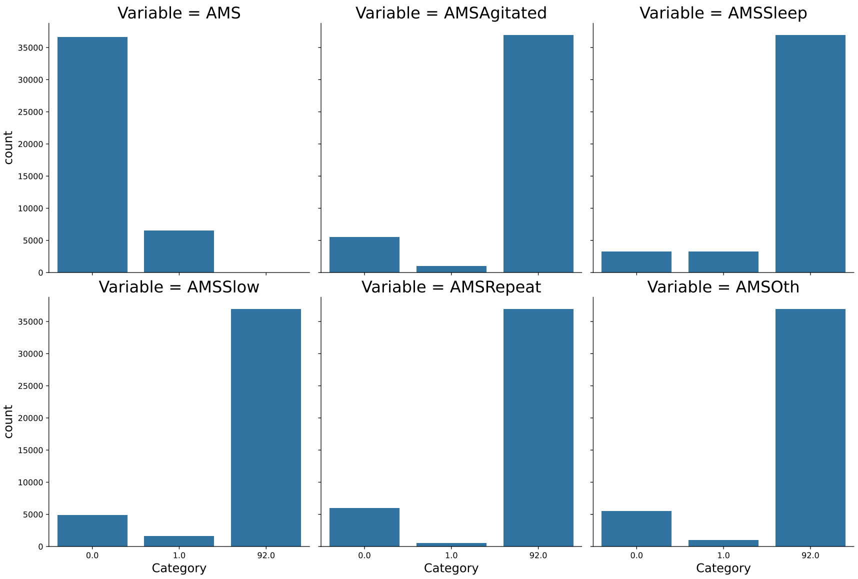

Figure 1: The figure illustrates the distribution of AMS related variables.

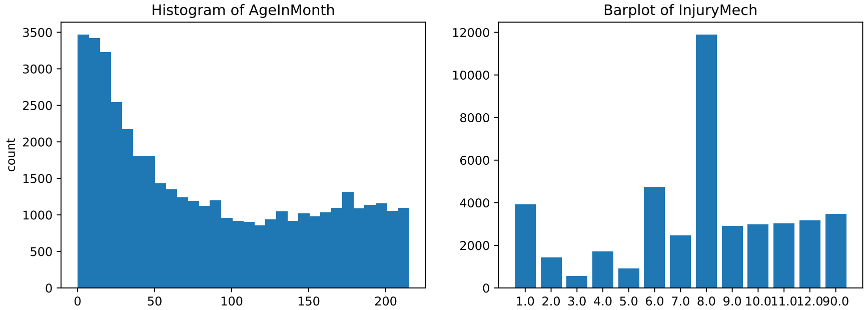

Figure 2: The figure illustrates the distribution of AgeInMonth and InjuryMech.

etc. represent different aspects. The distribution of them is shown Figure 1. The main numeric variable is AgeInMonth, and its distribution is shown Figure 3. Finally, the outcome variable is PosIntFinal. This is a highly imbalanced binary variable, indicating if the patient has ciTBI (Figure 1).

2.3 Data Cleaning

We are going to perform data cleaning following the instructions of VDS:

- Invalid or inconsistent values
- Improperly formatted missing values
- Nonstandard data format
- Messy column names
- Improper variable types
- Multicollinearity problem

2.3.1 Invalid or inconsistent values.

Since most variables are categorical variables and all the values have their explanation, there are no outliers observed. In this part, we will focus on the inconsistent data and errors in data.

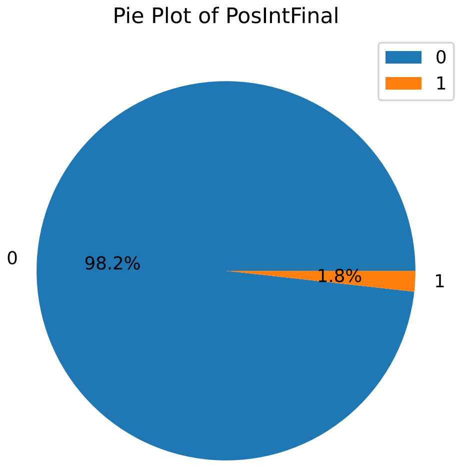

Figure 3: Pie plot of the outcome variable PosIntFinal.

First, we examine the numeric variables: Patnum, GCSTotal, AgeInMonth, and AgeinYear. Patnum is the identical index of every row of records and does not contain information of the patients or treatment. GCSTotal is the sum of GCSEye, GCSVerbal, and GCSMotor, which are different aspect of the GCS (Glasgow Coma Scale) score. We notice that although there are thousands of NA values in the three aspects, the GCSTotal scores are all clean. Besides, there are only 969 patients out of 42,412 with GCSTotal less than 14. We will exclude these samples since the risk of traumatic brain injury on CT is greater than 20% therefore the CT use is not controversial. After removing these samples, most of GCSEye, GCSVerbal, and GCSMotor scores are close to full credit and highly correlated with each other. Thus, we dropped the three columns and kept only GCSTotal for the following analysis. Finally, AgeInMonth and AgeinYear are both variables showing the age of patients, and we have checked that the numbers are matched. Since all the information of AgeinYear has been implicated in AgeInMonth and the latter is more detailed, we decided to drop AgeinYear to avoid duplication.

Then, we validated all the categorical variables. Here are the findings:

- HA_verb and Amnesia_verb. The two variables share the same ”not applicable” rule, which is ”patient is too young to speak of the patient is intubated or otherwise unable to give an understandable verbal response”. Therefore, they should be equal to 91 at the same time. However, there are several rows where only one of them equals 91 while the other one does not. In addition, there are 355 children less than 12 months old that were not marked as ”too young to speak” in the HA_verb, and so does Amnesia_verb. We are going to set the children less than 12 months old as ”too young to speak”, and align the values between the two variables.
- There are some cases where the physician indicated that the seizure is the most important indications in influencing the decision to obtain a head CT (IndSeiz=1), but Seiz=0, which means that the patient did not have seizure at all. Same things happen to Vomit, HA_verb, Hema, etc. There is an obvious inconsistency in these variables. For these problems, we assume the observation of patients is correct and the indicator variables are wrong for certain reason. We will modify the indicator variables to make them align with the observation. Specifically, we set the indicator variables to be 0 if the obervation equals 0 but the indicator does not.

Finally, we will drop all the unnecessary variables. Since the objective of the project is to identify the patients who do not need to take CT scan, it is impossible to take variables observed after CT or only valid for patients who took CT scan. These variables include 23 ”Findings” variables which are based on CT, 16 indicator variables (IndHA, etc.), EDDisposition, CTDone, EDCT, PosCT, CTForm1, DeathTBI, HospHead, HospHeadPosCT, Intub24Head, and Neurosurgery. Among these variables, even though HospHead and

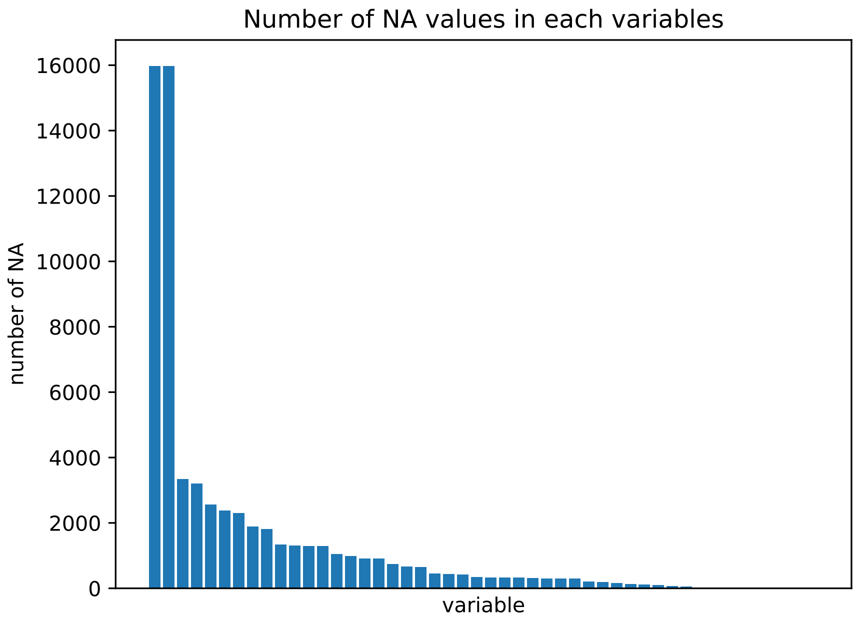

Figure 4: The figure illustrates the number of NA values in each columns decreasingly ordered by the number. For the clarity of the figure, we did not include the name of variables and did not show the variables containing no missing values. The left two columns represent Dizzy and Ethnicity.

PosIntFinal are highly correlated, we do not think it can be used as a predictor, because the correlation may come from the causality, i.e. the reason why patients has to be hospitalized is that there are some severe symptoms detected after taking CT. Therefore we are going to drop it.

2.3.2 Improperly formatted missing values.

The missing value problems are severe in this dataset. Most columns contain over 100 NA values, among which there are two variables containing significantly more NA according to Figure 4, Dizzy and Ethnicity. Since the two variables have over 35% missing values and the mechanism for which they are missing is still unknown, it is risky to include them in the predictive model.

In addition, we checked whether the missing data for each variable came from the same subset of samples. To achieve this, we used one-hot vectors to identify whether there is a NA value in the column and calculated the correlation coefficient matrix. We found that NA values are randomly assigned in most columns, and there are only several columns where NA values exist at the same time. These are all variables belonging to the same test, but representing different aspects. For example, GCSEye, GCSMotor, and GCSVerbal usually have NA at the same time. Therefore, it is not feasible to hugely reduece NA values by dropping only a few samples in the data set.

2.3.3 Nonstandard data format.

Variables in the data set are generally well formatted, but there is still a problem. We noticed that the values of AgeTwoPlus, Gender, Ethnicity are 1-2 instead of 1-0, which do not align with other variables and may cause some unnecessary problems while fitting models, making the model coefficients less interpretable.

2.3.4 Messy column names.

In the original dataset, the column names are not readable enough for the non-medical background researchers, and there are several errors or consistency in the format of column names (AgeInMonth and AgeinYear for example). Therefore, we are going to rename most of the column names based on the meanings. Here are some examples of new column names:

Old Names	New Names
InjuryMech	InjuryMechanism
Seiz	PostTraumaticSeizure
VomitLast	LastVomitingEpisode
AMS	AlteredMentalStatus
AMSRepeat	AMS_RepetitiveQuestions
SFxPalp	PalpableSkullFracture
SFxBasHem	BasilarFracture_Hemotympanum
Clav	TraumaAboveClavicle
OSI	OtherSubstantialInjury

Table 1: Comparasion between old and new column names

It has to be pointed out that these are not final names of variables, because we will transform some of categorical variables to dummy variables in the next part, and the column names will also change correspondingly, and we will add suffixes to identify dummy variables. For instance, AMS_RepetitiveQuestions may become AMS_RepetitiveQuestions_0.0, AMS_RepetitiveQuestions_1.0, or AMS_RepetitiveQuestions_92.0.

2.3.5 Improper variable types.

Since the data types of variables are all Int in the original data set, they are already clean enough and there is not much work need to be done. What we did is mainly to transform categorical variables to dummy variables to fit the need of predictive model. Considering the original paper has tried tree models, we decided to use other models to illustrate different insights, and this also increased the importance of the process of missing values and dummy variables. Specifically, for those variables with NA values, we regard NA as a kind of information and generate new dummy variables for them, using 0/1 as the indicator.

In addition, because we are not going to use tree models, the one-hot encoding of categorical features is necessary in our data cleaning process. For the same reason we mentioned in the 91 and 92 values part, after the one-hot encoding, there are 29 variables identical to other columns and we dropped them. The other details of cleaning process will be presented in section 2.2.6.

2.3.6 Multicollinearity problem.

The multicollinearity problem is significant in this data set, due to the several groups sharing similar meanings or scenarios. Since we plan to use generalized linear model to predict the outcome ciTBI, and the multicollinearity problem may lead to poor model interpretability and divergency, in this section we are going to handle this problem.

If we directly draw the correlation coefficient matrix plot as Figure 5, we can observe multiple blocks that are highly correlated with each other. However, we found that most collinearity arises from samples with values of 91 and 92. But after setting all 91 and 92 values as NA and draw the matrix plot again, we can see that the problem is significantly alleviated, as what is illustrated in Figure 6. This is because there are large amount of "not applicable" values and some columns have the same rules of "not applicable". For example, AMSAgitated, AMSSleep, AMSSlow, AMSRepeat, AMSOth are all marked as 92 if "patient does not have GCS < 15 or other signs of altered mental status or AMS is missing", which lead to the high correlationship. We think this does not shows the most essential informaton of variables and will only consider the coefficients after removing 91 and 92 values. Here are what we noticed:

- SeizLen, SFxBasHem, and NeuroDCranial are highly positive correlated, be this mainly becuase there are too many not applicable values in them. However, since the sparsity comes from the imbalance of the patients, we think they still contains valuable information, and will not drop the three columns.
- AgeInMonth and AgeinYears are positive correlated because thay represent the same meaning. We will drop AgeinYears.
- Observation variables and indicator variables like Seiz and IndSeiz, Vomit and IndVomit, are highly correlated. This is reasonable because physicians will mark most of the observed symptoms as the factors which lead to the CT examination. Additionally, since the indicator variables are only recorded

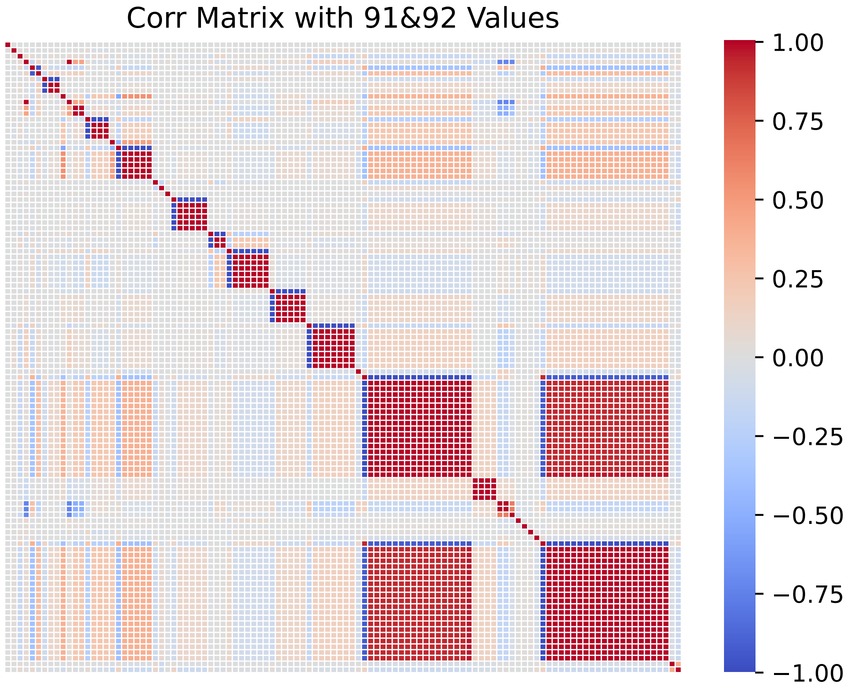

Figure 5: The figure is the visualization of the correlationship coefficient matrix. Every column and row represents a variable, and the color represents the correlation coefficient of the corresponding row and column. Since it is the overview of the relationship between variables rather than the specific values of coefficients that illustrates the point of this fiture, for the clarity of the plot, we did not include the column names in it.

when CT is taken, they are informative for the decision on whether someone need CT or not and we will drop these variables.

In addition to the correlation coefficients, after the one-hot encoding, we also used VIF scores to check multicollinearity problem. We noticed 22 variables with infinite VIF scores and 13 variables with VIF scores >10, indicating that they can be explained as the linear combination of other columns, and we will drop the 35 variables.

3 Findings

3.1 First finding

In order to discover the relationship between age and the outcome, we visualized the average trend of the relationship in the Figure 7. To reduce the variance of the estimation, we take the average of the outcome each 3 months. The blue line is the average outcomes in 3 months, and the purple band is the 95% confidence bound, the red line represents the spline estimation, showing the trends of the blue line, and the grey columns are the histogram of the AgeInMonth variables which is to show the distribution of Age and confidence of the estimation. Based on this figure, we can see that there is a shift of then trend around 12 months. When AgeInMonth<12, there is a decreasing trend, and it increases after 12. The red dash line markes 12 months. Based on this finding, we added two variables, AgeLessOne and AgeLessOne_inter.

3.2 Second finding

We drew the Figure 8 to demonstrate average outcome PosIntFinal under different InjuryMechanism. There are 4 mechanisms have significantly higher risk comparing to other ones. The 2, 3, 5, 1 represent ”Pedestrian

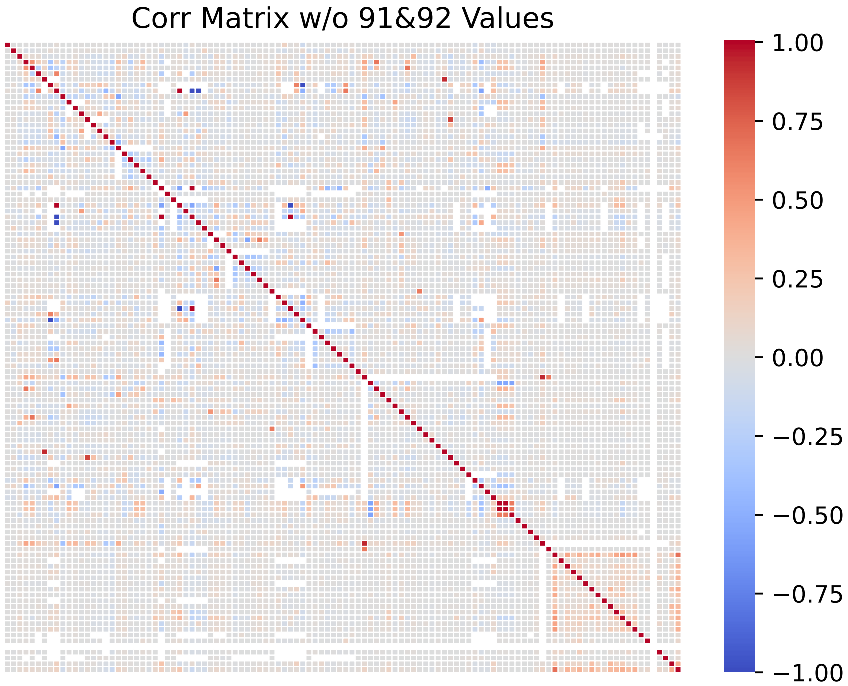

Figure 6: The figure is the visualization of the correlationship coefficient matrix after setting all 91 and 92 values as NA.

struck by moving vehicle", "Bike rider struck by automobile", "Other wheeled transport crash", and "Occupant in motor vehicle collision (MVC)". This shows that comparatively, traffic collisions are the primary factor of a high risk of ciTBI, with pedestrians being at the highest risk, followed by cyclists, and then automobiles, which makes sense because pedestrians normally have the least protection.

3.3 Third finding

Similar to the finding 2, Figure 9 shows how clavicles evidence influences the outcome. We removed samples without any evidence of trauma. It shows that the influence of the injured area on the presence of ciTBI is as follows: face > scalp-frontal > scalp-occipital > scalp-parietal > scalp-temporal > neck.

3.4 Reality Check

- Do a reality check. What reality could you compare your cleaned data to?
- Clearly state your assumptions and explain why this reality check is useful.
- Does your cleaned data pass the reality check or are there issues? Discuss.

For the reality check, we did the following items:

- Check whether there are NA values in the data set.
- If the distributions of variables are reasonable. For this part, since we did not drop a large amount of samples, the distributions almost remain the same as the original data. The main potential problem is the imbalance problems of several variables, but in reality the patients with ciTBI is the minority according to the original paper, so we think this still makes sense.
- The second problem is the inconsistency of values as what we mentioned in Section2.3. We checked the data after data cleaning, and did not discover inconsistent values again.
- Finally, we regard the distribution of variables in the original paper as one of the standards, and

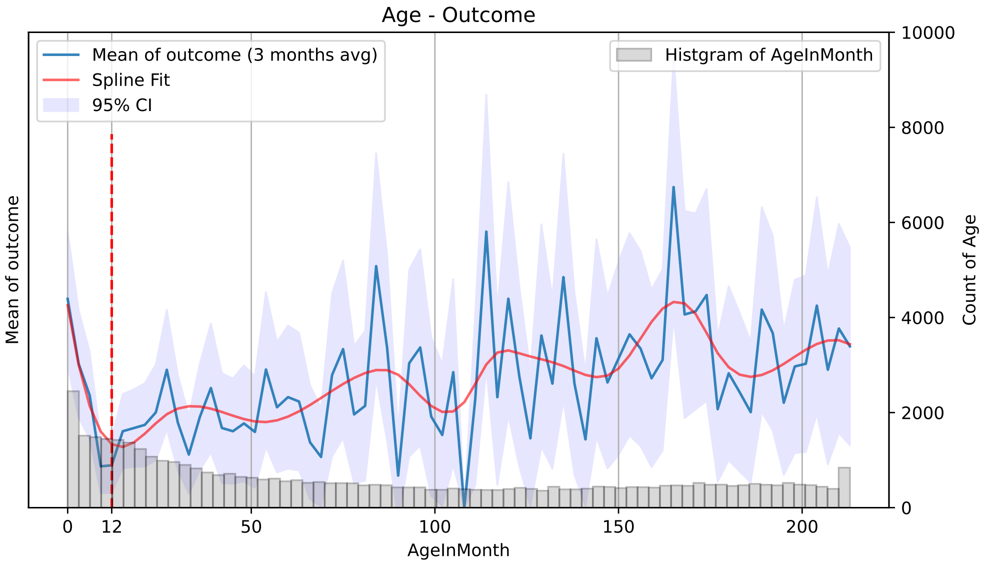

Figure 7: The figure demonstrates the relationship between AgeInMonth and PosIntFinal. The blue line is the average outcomes in 3 months, and the purple band is the 95% confidence bound, the red line represents the spline estimation, and the grey columns are the histogram of the AgeInMonth variables

compared proportion of different values. We did not notice significant deviations between the two distributions.

Based on these analysis, we think the data cleaning has passed the reality check, and we are going to discuss more about modeling in the following sections.

3.5 Stability Check

Take one of your findings and present a perturbed version. How does this affect your finding? Add a before and after plot here.

To test the stability of the findings, we took Finding 2 as an example. Since the PosIntFinal is a binary variable, it is infeasible to add gaussian noise to it. Thus we randomly chose 30% samples and shuffled their PosIntFinal to introduce noise to the data set. Then we drew the plot again to compare the difference (see Figure 10). Overall, the values of high risk group are still significantly higher than the other variables, and the rank did not change except for value 11 and 9. This proves that finding 2 is stable.

4 Modeling

4.1 Implementation

For the first model, to handle this binary classification problem, considering the model interpretability and stability, we are going to use generalized linear model. Specifically, we used logistics regression with L1 regulation to get a sparse result.

$$\text{logit}(P(y=0|X=x)) = x^\top \beta,$$
$$\mathcal{L} = \mathcal{L}_{CE}(\hat{y}, y) + \lambda \|\beta\|_1,$$

where $\mathcal{L}_{CE}$ is the cross entropy loss. In addition, since for this project, we hope to control the false negative rate in order to discover more potential patients and avoid treatment delay. Thus, we assigned more weight to positive samples, and the loss function becomes:

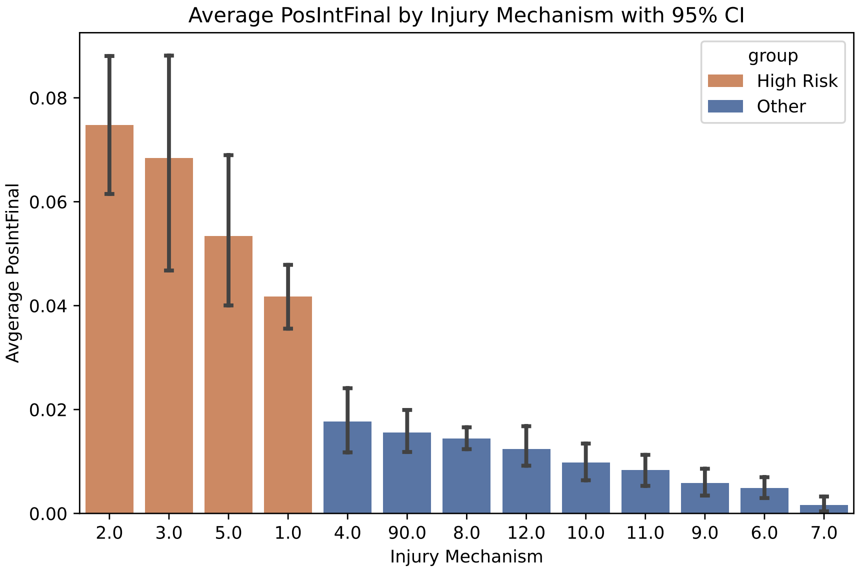

Figure 8: The figure demonstrates average outcome PosIntFinal under different Injury Mechanism, ordered by the average values. The plot also shows the 95% confidence interval.

$$\mathcal{L} = \frac{1}{N}[\sum_{i\in\{i:y_i=1\}} w_1\mathcal{L}_{CE}(\hat{y}_i, y_i) + \sum_{i\in\{i:y_i=0\}} w_0\mathcal{L}_{CE}(\hat{y}_i, y_i)] + \lambda\|\beta\|_1,$$

where N is the sample size, w_0, w_1 are the weights and we set $w_0 = 0, w_1 = 100$.

To determine the hyperparameter λ, we used the cross validation, where we split the data set into training, validation, and test sets in a 6:2:2 ratio and use grid search to select the λ that performs best on the validation set. As for the evaluation metric, since the labels are imbalanced in this data set, the positive samples are obviously less than negative ones, we will use AUC rather than accuracy as the metric.

As the result, there are 26 variables with nonzero coefficients, the AUC on test set is 0.94 with the hyperparameter $C = 0.03$, and the confusion matrix is shown in Figure 11.

Similarly, for the second model, we chose SVM, and keep all the other settings including training-test split, seed, weight, and hyperparameters choice (cross validation) the same, but without L1 regularization due to the algorithm of SVM models. The AUC on test set is 0.93 (with the hyperparameter $C = 0.1$), and the confusion matrix is also shown in Figure 11. It should be minded that, if we do not specifically add weights to positive samples, it is easily to get a naive result where all the predictions are negative.

Overall, the performance of the logistics model is better than the SVM model, both AUC and FNR of logistics regression are all better. Besides, the interpretability of the two models are different, and we will talk about this in the section 4.2.

4.2 Interpretability

For the two models we used in this project, the logistics model has more interpretability than the SVM model does. Unlike tree models, it is hard to have a clear visualization of the decision boundary of these models if the dimensionality of X is larger than two, so we did not include the figure to show it.

The reasons why logistics model is more interpretable are that:

- The decision boundary of the model is linear, and through the coefficients we can tell how different variables contribute to the prediction.

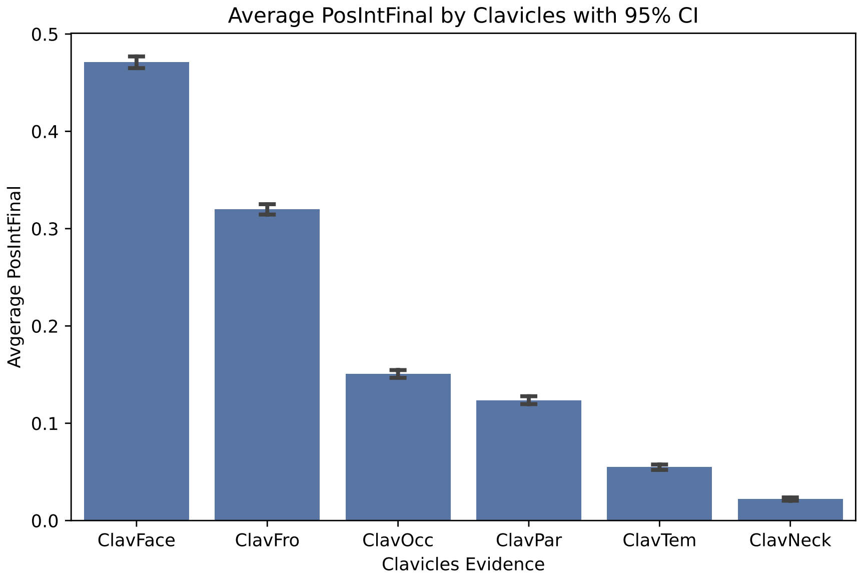

Figure 9: The figure demonstrates average outcome PosIntFinal under different clavicles evidence, ordered by the average values. The plot also shows the 95% confidence interval.

- We have a sparse model due to the L1 regularization, making the data collection and decision making process much easier obtains the predictions it does?

The SVM model requires complete variables and the decision boundary in the kernel space remains unknown, which makes it harder to explain to the healthcare providers. Since the SVM model has poor interpretability, if we care more about the accuracy rather than the model interpretability, it may be more feasible to use more complex models like random forest thay may have better performance.

4.3 Stability

To check the stability of two models, we shuffle the outcome with different proportion and check the AUC values on the test set. The value is shown in Figure 12. Overall, the AUC remains stable when shuffle rate is less than 30%, showing the stability of two models. Specifically, we shuffle the outcome variable in the training and validation set, and keep the test set constant, by doing this, we can test the model performance while there is noise in the training data.

5 Discussion

- The data size is relatively large, with more than 40,000 samples, making it possible to train basic machine learning models. But it restricts the computation resources since the larger data set brings longer training time. For the two models we used, it took much longer to train SVM models especially when the hyperparameter is small. Therefore, the simulation or Monte Carlo experiments is less feasible for this algorithm.
- The main problem of data / reality is that, the objective of the project is to find a rule to determine whether a patient should take CT scan. However, excessive variables limited the feasibility, because it is hard to collect this many variables for every patient, which also demonstrate the significance of the sparse models. Therefore, we think the logistics model is more feasible for the application, and the tree models used in the original paper also work well. As for the future data, it is reasonable to assume the distribution of variables remains the same in a period of time, like 5 years. This also depends on how the data is collected, including whether there is selection bias in the data collection process.
- In this lab, the EDA, data cleaning part is related to the data / reality. The model training fits

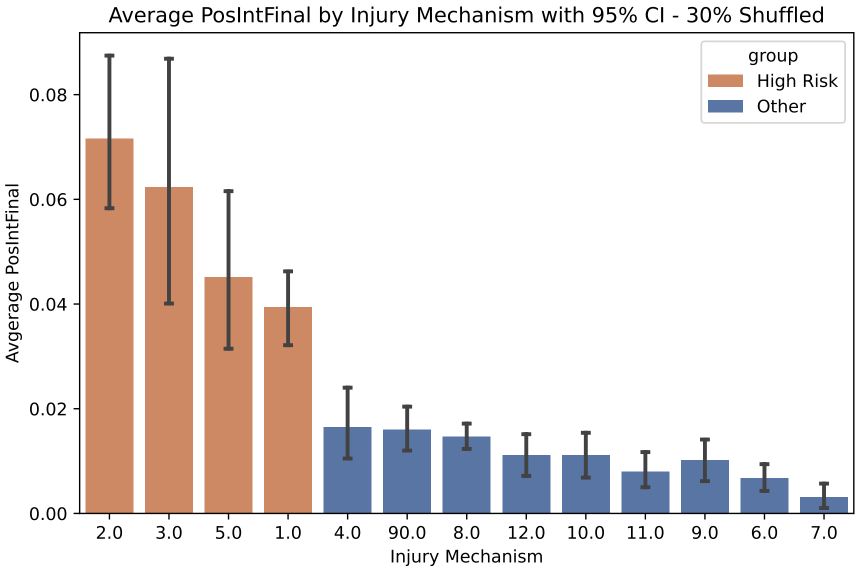

Figure 10: The figure demonstrates average outcome PosIntFinal under different Injury Mechanism, ordered by the average values, but with 30% PosIntFinal being shuffled.

algorithms / models. The discuss about the stability and interpretability fits the future data / reality.

- The reality is definitely much more complex than the data. Because there are much more factors affecting the outcome, and the collection process also introduces noise, including selection bias, human mistakes, etc. But what we can do is to make full use of the data that we have access to.
- The data visualization is also hugely influenced by how we clean it. For instance, whether we regard the 91/92 values as NA, whether we delete all the NA values, how we decide which rows or columns to drop also have impact on the visualization, leading to different results.

6 Conclusion

In this work, we discovered analyzed the data set to build a prediction model, and provided advices for the medical personnel based on the three findings. First, besides the decision tree in the original paper, we showed that the logistics regression with L1 regularization also works well. Comparing to SVM models, the logistics regression has better interpretability, as well as better performance. Based on the three findings, we recommend giving extra attention to patients with facial trauma and scalp-frontal injuries, as well as those with insufficient protective measures in traffic accidents. Additionally, we found that for infants under one year of age, the risk of ciTBI increases as age decreases, while for individuals older than one year, the risk tends to increase with age.

7 Academic honesty statement

Dear Bin,

I affirm that all the work in this report is done on my own, and all sources, including classmates, are properly cited. For me, academic honesty is essential and it is the responsibility of all students. However, I do not think the honesty statement is necessary for every report. It is more of a formality rather than a constraint.

Xuanlin Mao

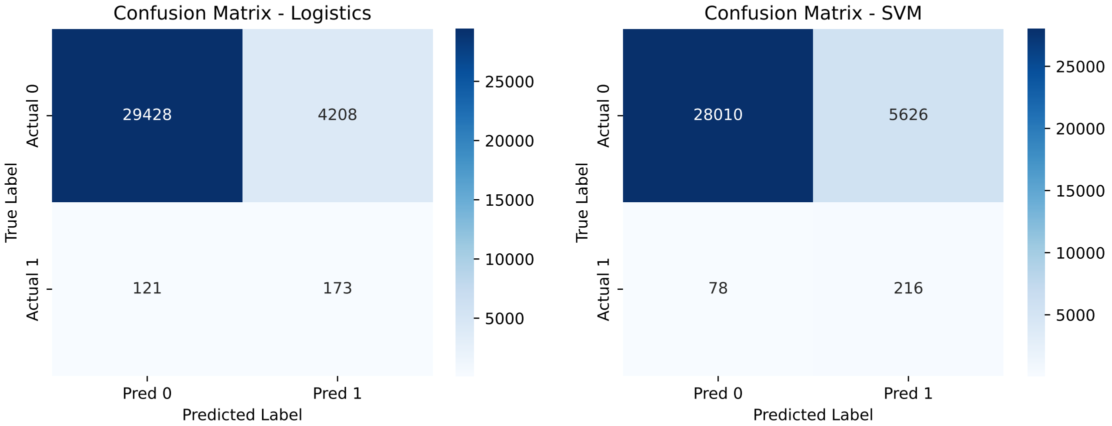

Figure 11: The figure shows the confusion matrix of model performance on test data set.

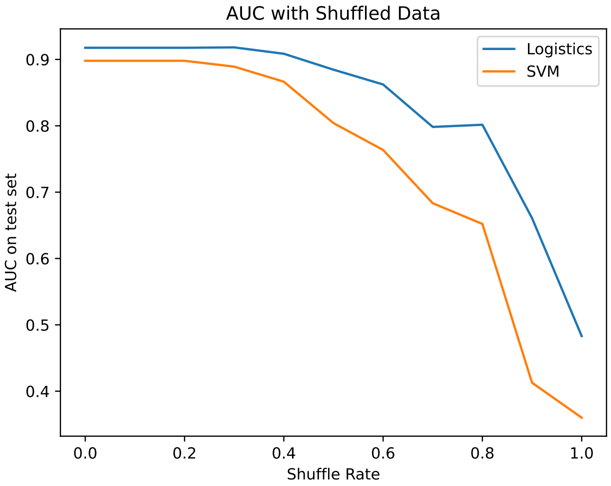

Figure 12: The figure illustrates the change of AUC of two models while shuffling the data set.

8 References

[1] Kuppermann, Nathan, et al. ”Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study.” The Lancet 374.9696 (2009): 1160-1170.