STAT 215A, Final Project: Classifying ciTBI in Youth

Mark Oussoren, Sahil Saxena, Hyunsuk Kim, Florica Constantine

12/10/2021

1 Introduction

Traumatic brain injuries, hereafter referred to as TBIs, are both commonplace and require immediate medical attention. However, diagnoses often require a CT scan to confirm the presence of a TBI [CITE 14-16 from paper]. In children, this need is problematic, as the radiation from a CT scan can lead to long-term adverse affects; hence, given a child presenting with a potential TBI, it is desirable to find a way to decide whether they actually need a CT scan. Note that forgoing a CT scan in the presence of an actual TBI is also undesirable. In this report, we revisit the data from [CITE]. That is, we derive an updated decision rule for identifying which child patients need a CT scan.

TODO: OUTLINE OF SECTIONS

2 Data

2.1 Data collection

The authors in [CITE] collected data in a prospective cohort study from patients younger than 18 years of age that visited a hospital within 24 hours of experiencing head trauma. The study was run across 25 pediatric emergency departments over a span of approximately 2 years, where the last few months were used to collect samples for validating the decision rules derived in the original study. Only patients with GCS scores of 14 or 15 were considered; those with scores 13 or less were enrolled but were not grouped with the others. For each patient, a trained investigator or other medical personnel recorded various prespecified details, e.g., mechanism of injury, medical history, and responses to standardized questions about the presence of several symptoms or signs of head trauma. For a small subset of patients (approximately 4%), a second assessment was performed for quality control purposes–note that we do not use this information, but that its presence is reassuring.

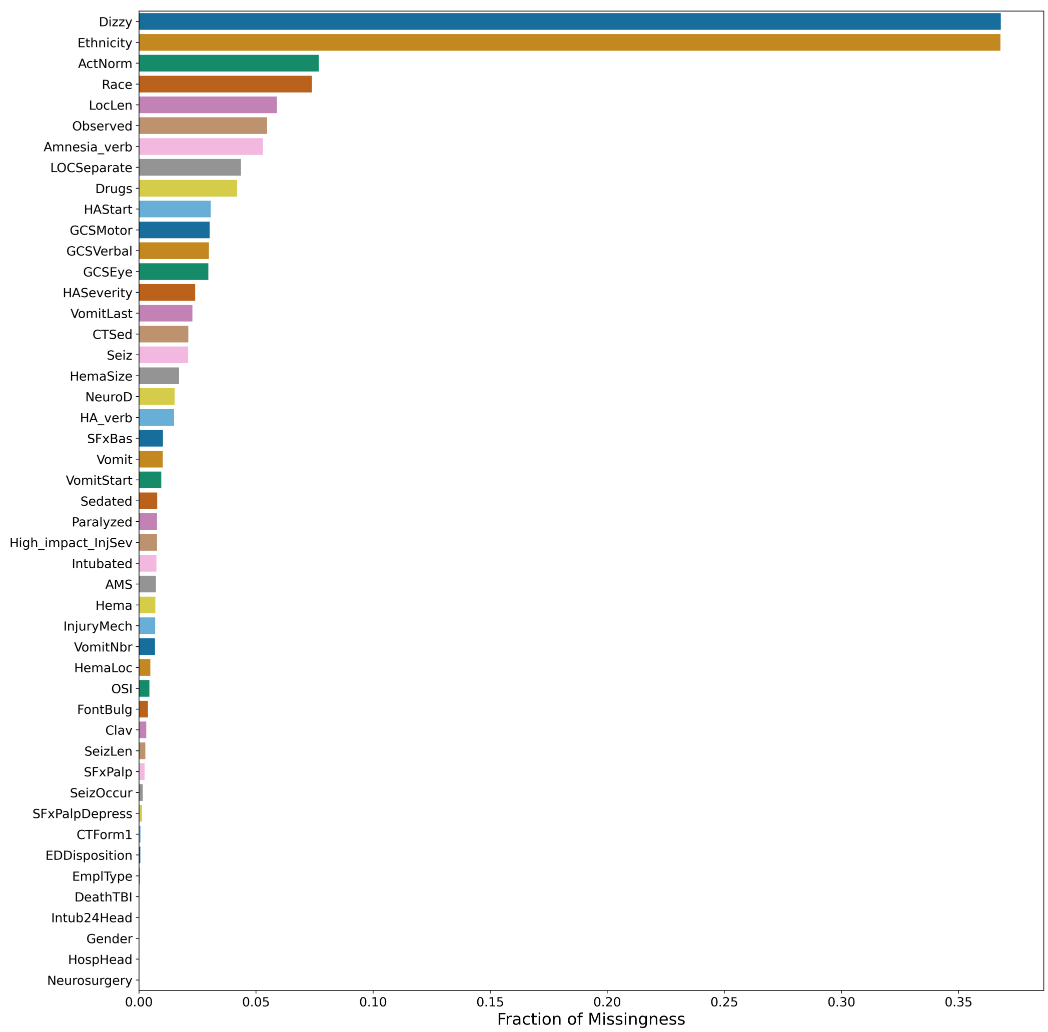
The study defined death, prolonged hospital admission or intubation, or the need for surgery following a CT scan as a postitive, ciTBI outcome. Other patients were assigned to the negative outcome; to find missed positives, the study coordinators performed telephone surveys to follow up with parents and tracked followup visits. If a positive outcome was missed, the patient's label was updated to positive.

Note that several variables or descriptors in the study require the ability to converse with the child for assignment, e.g., the presence of a headache. Hence, the authors chose to separate patients under the age of two (pre-verbal) from those aged two or older (verbal) in their analysis. Moreover, as children under the age of two are more sensitive to radiation, it is reasonable to consider this group separately.

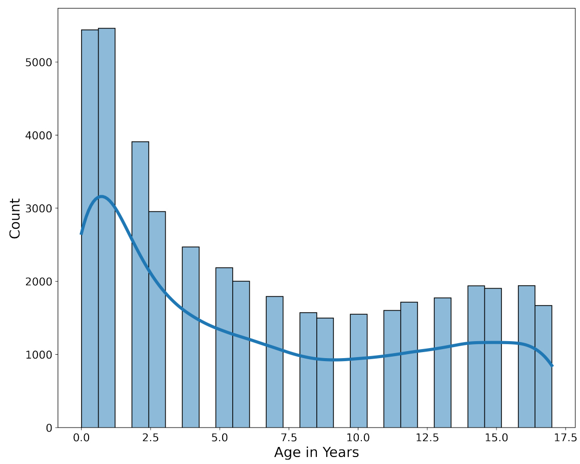
2.2 Meaning

2.3 Exploratory Data Analysis

Our data set initially consists of 43,399 patients under the age of 18 with mild head trauma evaluated in 25 PECARN emergency departments. In particular, we have 125 features filled out by trained site investigators and other emergency department physicians who recorded patient history, injury mechanism, and symptoms and signs on a standardized data form before knowing results along with the results recorded as 'PosIntTBI'. We first noticed that the data set had several variables that were not filled out - recorded as nans or np.nans. Below, we can see the fraction of missing data for each of the features. Notably, there are features "Dizzy" and "Ethnicity" that are missing in more than 35% of the points. Looking through the variable descriptions and speaking with the clinicians tells us that these are in fact not very relevant and very susceptible to change patient by patient. For these reasons, we end up dropping these variables.

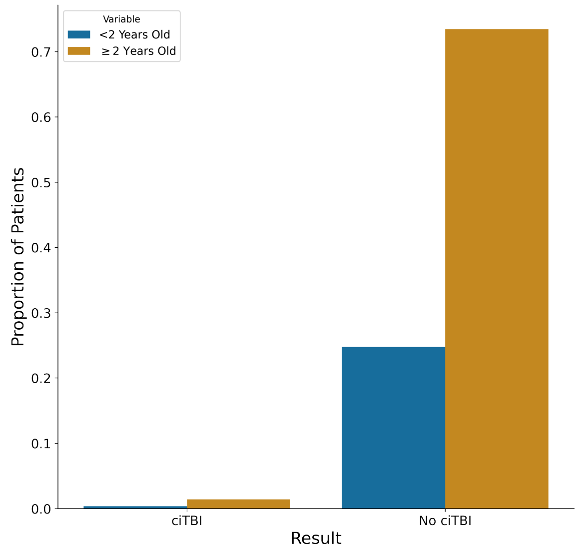


Next, as reality checks, we notice that there are 20 patients with the variable 'PosIntFinal', our desired outcome, missing however in the paper however, they claim only 18 are missing. This discrepency is noted, but not much else can be done. However in the related outcome variables which consist of "Intub24Head", "Neurosurgery", "HospHeadPosCT", and "DeathTBI", their union of missing/unknowns is only one which means at most one of the 20 outcomes cannot be inferred indirectly through these. As seen below, it turns out that we can infer all of these outcomes which results in no missing outcomes from our data set. Outside of this difference between our data and the paper, we can compare the GCSTotal scores. Notably we find that there 969 GCS scores that lie between 3 and 13 as described in the paper. As an aside from the paper comparison, we notice that the main continuous variable in our repertoire is the age. The age of patients from this dataset is plotted below, and we notice a very significant chunk of patients are two years old or younger. As seen in the paper, this split in frequency is also marked by not so similar distributions of the outcome hence why the paper partitions the dataset into two groups.



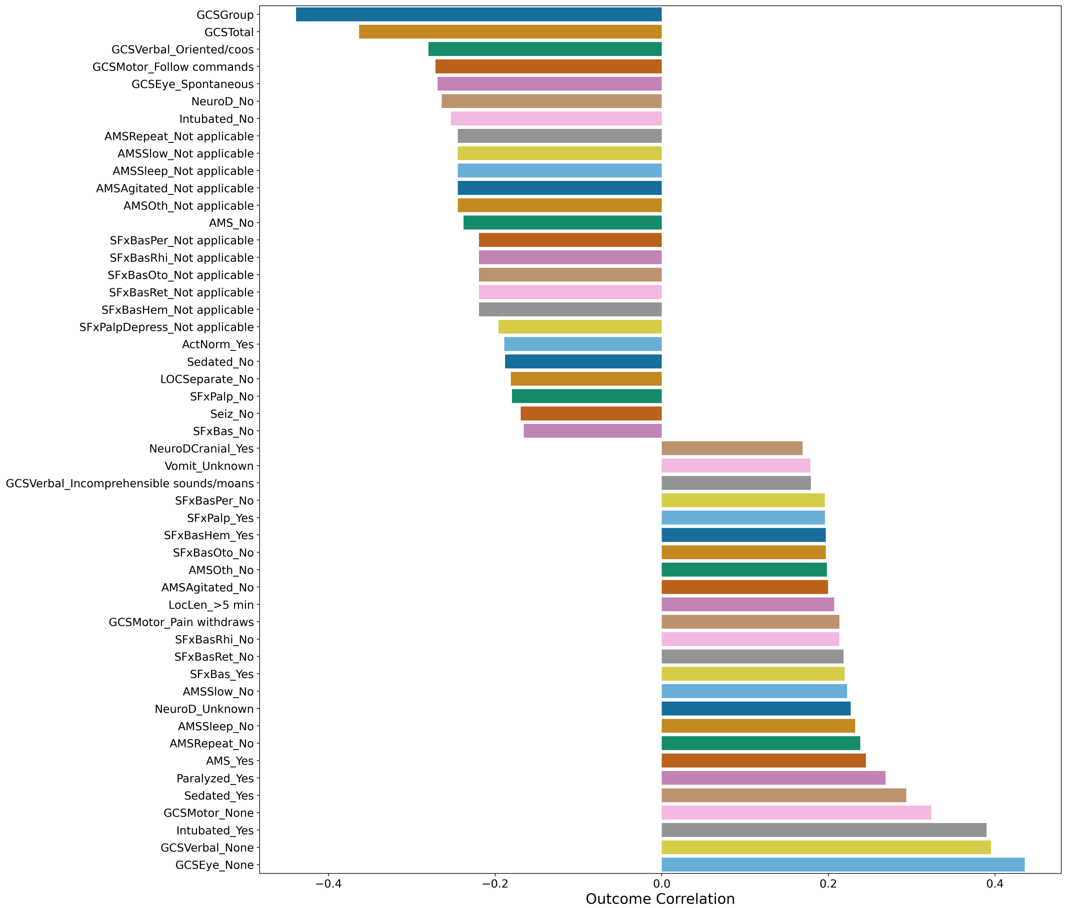
To retrace back to our analysis of missing features, we would now like to shine some light on our judgement calls and the cleaning of the data.

For starters, we sifted through the excel sheet containing variable names and descriptions and noticed that many variables list the category "Not applicable" as "No" or "Missing". To determine which variables exactly we could impute 'Not applicable' as "No", I simply match up those "Not applicable" ones that correspond with "No" in the parent question. For instance, we have "Seiz" which indicates if the patient had a seizure or not and many subsequent questions with "Not applicable" as a response. We find all of the "child" variables can be imputed as no because they all correspond to "No" in the "parent" variable. Next, we turn to impute the missing entries discussed earlier. These make up a very small percentage of the data as seen in the bar plot above. For this reason, we initially decided our judgement call would be to fill in the features not described in the data description excel file with the most likely or mode. *Do I need more justification for this - it's a very small subset of our data?* On top of this, we decide unlike the paper, that we will not necessarily fit our model on two separate age groups. After looking at the distribution of outcomes in the two groups, the difference is not very noticeable. However, when comparing our models to the baseline, we will in fact create a fair comparison by training on the age splits referenced in the paper.

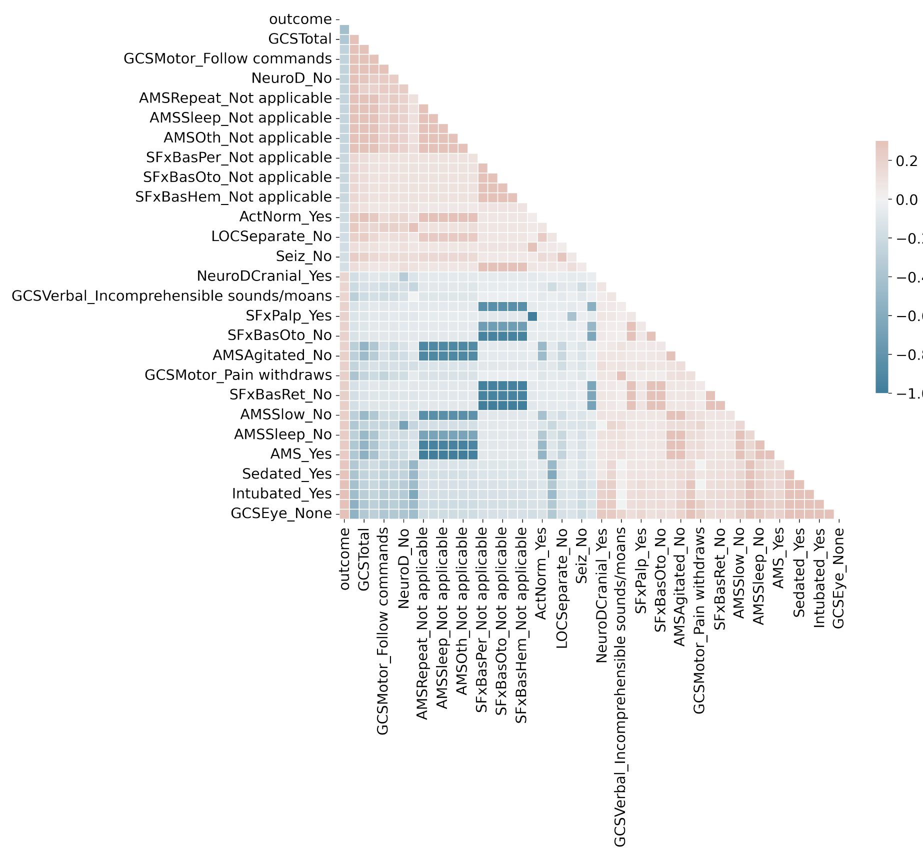


From this cleaning, we have generated two data sets. The first we consider is what we shall call the simple data set. This consists of 42430 with 44 columns (including the outcome - if a patient has ciTBI or not), 0 missing entries, and all variables are binary (yes / no) as GCS scores and age were removed and other categorical variables were one-hot-encoded. The second data set we consider is if we only remove the variables that occur after CT decisions are made, those with no importance in the decision (ethnicity, race, etc.), and again one-hot encode the variables. This data set consists of 42430 patients with 120 columns (again GCS and age variables were removed with only one-hot encoded variables remaining).

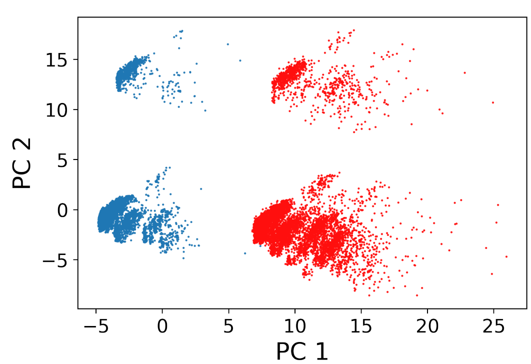
Finally, we pivot to discovering variables that may prove fruitful in determining whether or not a kid has ciTBI. First, we may simply analyze the features that are most correlated in absolute value to the outcome after one-hot encoding all of the variables and looking at Spearman's rho. Notably, all of the GCS (total, group, and sub rankings) alongside AMS ans SFxBas variables (and sub variables as these include multiple sub parts) are the most correlated with the outcome compared to the other features.



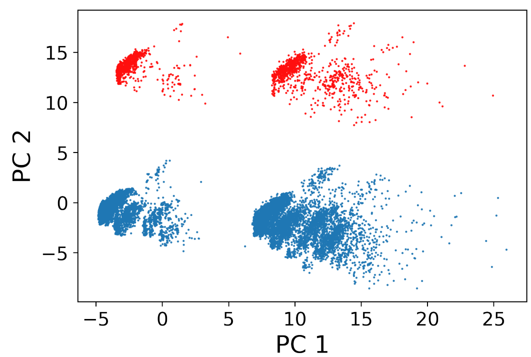
Interestingly enough from the plot, if we remember from our earlier analysis of missing variables, the GCS sub features are more missing than the total and they both share very strong correlations to the outcome. To reconcile which are more important (as this is necessary in several classification models such as logistic regression), we plot the correlations of these correlated variables in the heat map below. From this analysis, it is clear that the GCS variables are correlated as indicative by the blues, however what is even more correlated is the string of AMS variables. Again, as discussed earlier, the questions with nested/follow ups move hand in hand with each other. This line of reasoning forms the basis for our initial data set for modeling - strip the data to only consist of those parent questions along with those that have no follow up questions but are still relevant.



Finally, to close our EDA we can pivot to a more sophisticated method for explaining our data - through PCA. In the first PC plot, we notice CTForm1 which determines if a CT was ordered or not contributes the most to PC1. In hindsight, this will obviously separate our outcome as it is typically decided after the decision is made. Thus, this variable along with several other will be removed for the sake of not biasing our rules.

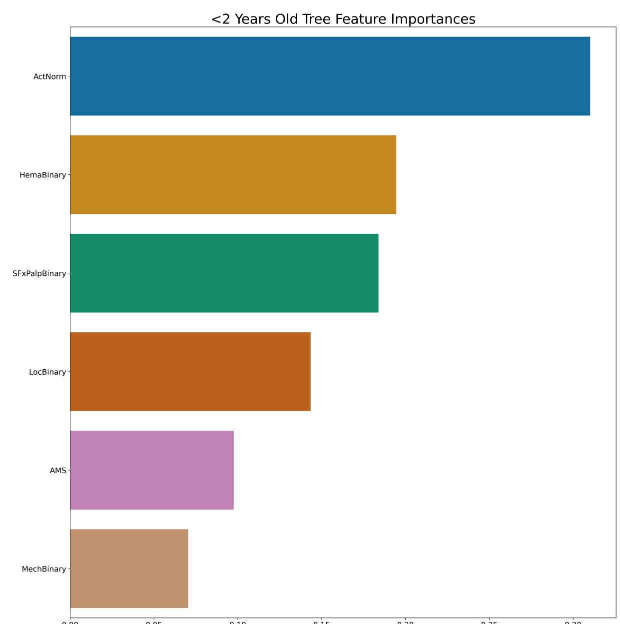


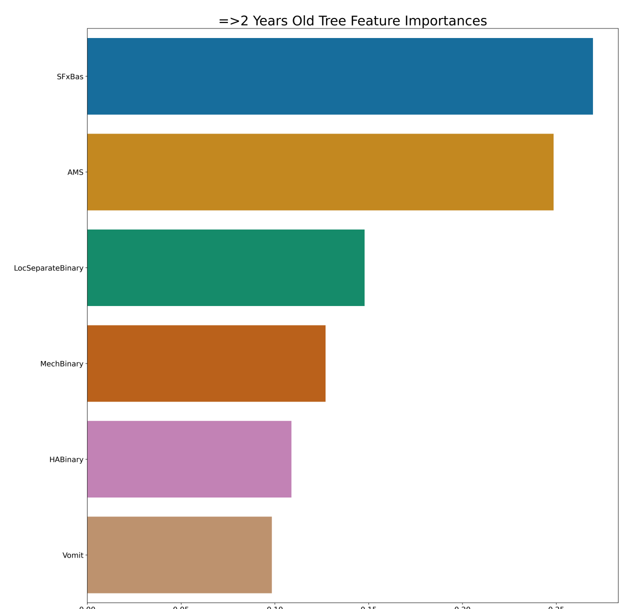
In this latter PC plot, the second PC whose largest loading is OSI, which determines if another non-head substantial injury occurred. After speaking with the clinician, we decided that this variable while potentially giving of the outcome is most likely pre-CT and can thus be integrated into our modeling as it separates the outcome (blue versus red) quite nicely.



3 Baseline Model

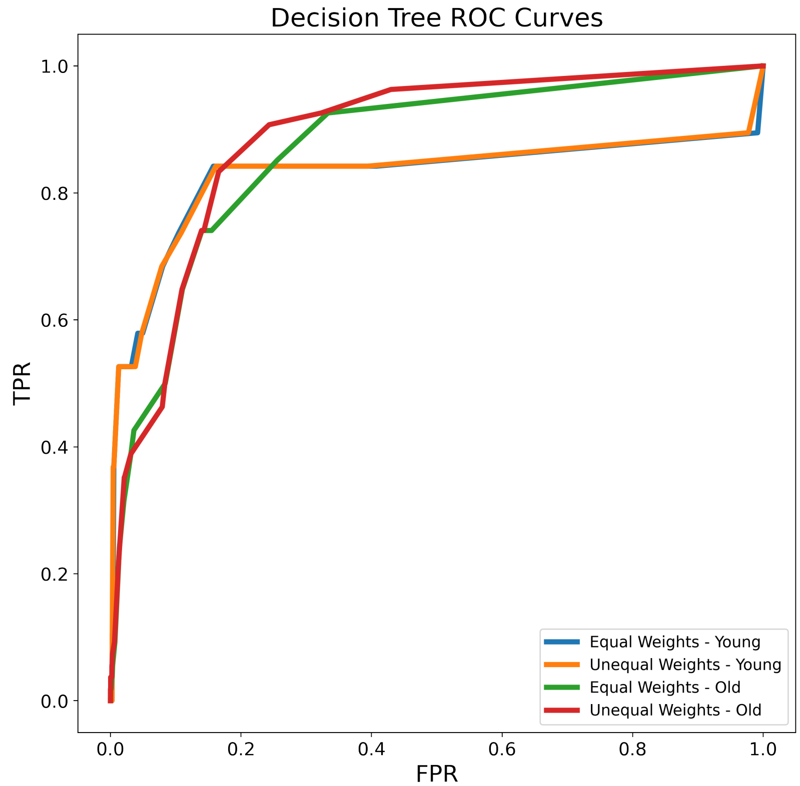
In this section, we will discuss how exactly we decided to reformulate the baseline model as described in the paper. As discussed in the EDA section, we ensured that we were starting with the same data - as noted we found 20 missing outcomes instead of 18, but aside from this we found the data to match up nicely. To further replicate the data used, we noticed the paper transformed several variables to be binary - 'HemaLoc', 'LocLen', 'High_impact_InjSev', 'HASeverity', 'SeizLen', 'HemaSeize', 'LOCSeparate', and 'SFxPalp'. The paper binned several categories of these features together into 'Yes' and 'No' which they describe in the decision tree visuals created. After making this change, we partition the data into two groups - those with kids less than two years old and one for those that are greater than or equal to two. From here, to mimic the same results we use the same features derived in the paper alongside the same depth and class imbalances. Oddly enough, they cite weights of 500:1 "for failure to identify a patient with ciTBI versus incorrect classification of a patient without ciTBI". The intuition for this weighting scheme is due to the imbalances in the classes to begin with however the exact cost is rather arbitrary. Anyhow, with this in mind, we fit the two decision rules - one for young and one for old kids. The plot for children less than two is interesting as it does not match what the paper found whatsoever in terms of the order of important features while the second tree shows comparable important features. As far as scores are concerned, we obtain the following metrics (young model, old model) - AUC: (0.855, 0.853), Accuracy: (0.854, 0.853), Sensitivity: (0.855, 0.833), Specificity: (0.789, 0.796), and Balanced Accuracy: (0.822, 0.814).





If we instead consider the weightings of 500:1 instead of 1:1 for the labels, then as below, we can see the

ROC curve do not have meaningful shifts in AUC figures.



However instead of trying to replicate the models fitting our own models with no certainty about their hyperparameters, we can simply traverse through the cleaned data set in the exact same fashion as they did and obtain the probabilities for each node. We carry out this and the code is embedded in the Python file and Jupyter notebook.

4 Modeling

4.1 Decision Trees

asdf.

4.2 Logistic regression

asdf.

4.3 Boosted Models (Ada & ..)

asdf.

4.4 SVM (and Kernel SVM)

asdf.

5 Conclusions

asdf.

5.1 Division of Labor

asdf.

6 References