Predicting Postoperative Wound Infection in Pediatric Plastic Surgery Patients

By: Nick Sawicki

Introduction:

Postoperative wound infection in pediatric patients receiving plastic surgery is an important area of study due to the adverse outcomes that can result because of infection. From an acute standpoint, patients with postoperative wound infection are at risk for developing more serious systemic infections if the initial infection is left untreated and spreads to the rest of the body. This prolongs hospital stays and puts patients lives at risk in the process. From a long-term standpoint, postoperative wound infections can lead to suboptimal aesthetic outcomes in patients which leads to future hospital admissions, increased risk of developing future infections, and potential psychosocial damage.

Thus, using the American College of Surgeons (ACS) National Surgical Quality Improvement Program's (NSQIP) Participant Use Data File (PUF) for Pediatric Patients, my aim is to develop a machine learning model that can predict whether or not a pediatric patient receiving reconstructive plastic surgery is at risk for developing a postoperative wound infection.

Data from NSQIP's PUF file is collected from participating hospitals across the United States by Surgical Clinical Reviewer (SCR) who are trained and recertified annually by NSQIP. Audits of each hospital are routinely conducted to validate the integrity of the data collection process, with an average discrepancy rate of 2%. ACS maintains a 5% threshold for acceptable disagreement rates, and hospitals surpassing this threshold are ineligible to receive hospital odds ratios/must undergo additional auditing/recertification.

The data set pertaining to pediatric plastic surgeries contains data from 55,432 pediatric patients, and collects data on 370 different variables. Data can be procedure agnostic, such as age, race, and sex, or highly dependent on procedures, such as time spent in operating room, whether or not the patient was receiving chemo therapy, or if they went into cardiac arrest. Upon requesting access to the data set, a user guide was provided describing all of the possible options for the various categorical and ordinal features. The target variable in this particular study is whether or not a patient developed a postoperative wound infection – a categorical yes/no.

Exploratory Data Analysis:

Due to the large size of the dataset, both with respect to the number of patients recorded and the number of features included, I decided to create a new dataset by randomly undersampling the large one to perfectly balance out the target variable. Originally, out of the 55,432 pediatric patients that data was collected on between the years 2012 and 2020, 807 reported postoperative wound infections less than 30 days after surgery was performed. This is a highly skewed distribution for the target variable, with balances of 1.45% in the 'yes' class and 98.55% of patients falling into the 'no' class for postoperative wound infections. By randomly undersampling the dataset, I was able to generate a perfectly balanced dataset of 1,614 patients where 50% or 807 reported postoperative wound infections 30 days after surgery and the other 50% reported no postoperative wound infection. Though undersampling does not come without its risks of excluding potentially relevant information from model and hyperparameter training, it ultimately allowed models to be trained more efficiently, while also mitigating the potential tendencies to overpredict class 0 (no postoperative wound infection) purely because of its relatively elevated prevalence.

In image 1 shown below, when plotted against the procedural category, it's interesting to note that there was an imbalance in wound infection prevalence between the different categories. General Reconstruction procedures were more likely to have patients with postoperative wound infections while Breast, Craniosynostosis, and Hand and Peripheral Nerve procedures showed almost no postoperative wound infection.

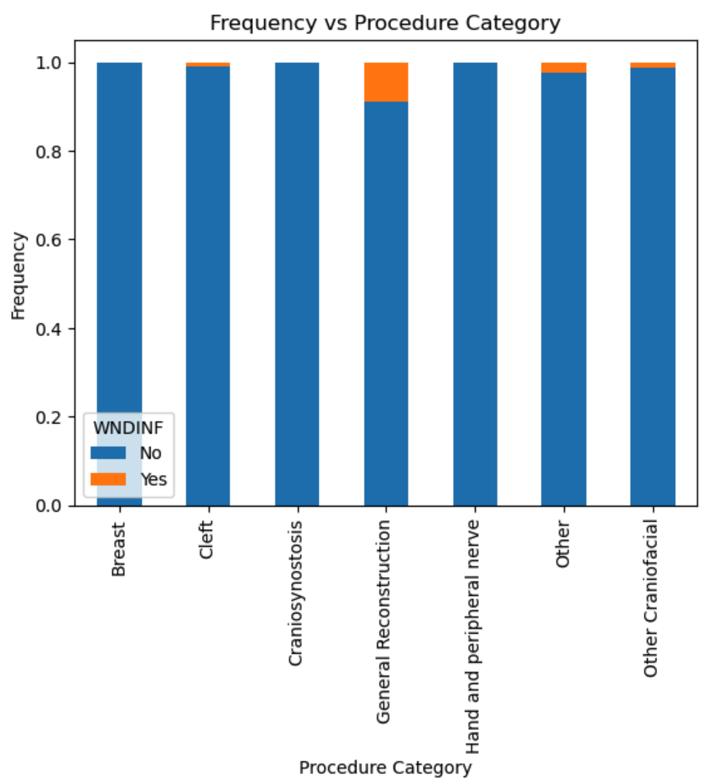

Image 1: The normalized plot above above plots procedure category on the x-axis and prevalence of wound infection on the y-axis.

In the box plot shown in image 2 below, age (measured in days) also seems to predict postoperative wound infection, with older children more likely than younger to develop infection.

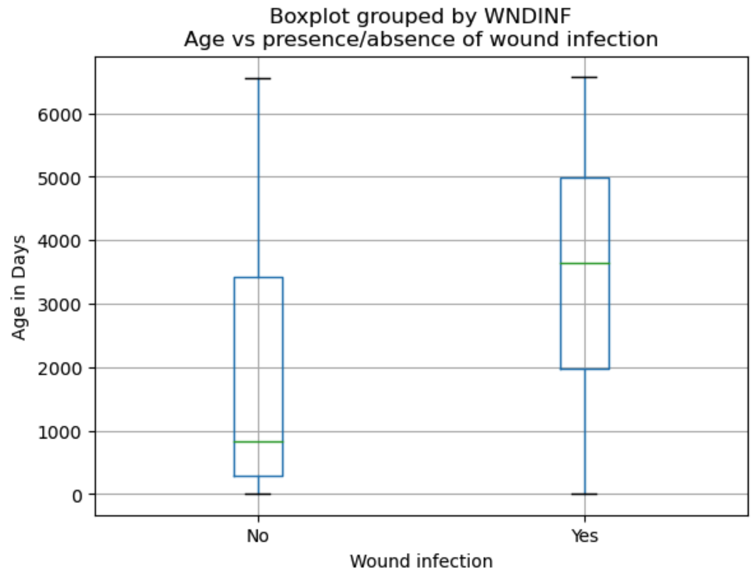

Image 2: The box plot shown above plots age in days on the y-axis and wound infection status on the x-axis.

Many of the 370 features included in the original data set included no actual values – they were completely devoid of any information. Therefore, the decision was made to eliminate features that contained 80% or more null values. This reduced the number of features down to 81. Four additional continuous features were eliminated due to the fact that they were missing over 50% of their values. Due to the fact that this is a medical data set, I did not feel comfortable imputing that many missing values and would lose significant amounts of data by dropping individual rows with missing continuous features. Features with correlations of 1 and -1 were also eliminated from the dataset in order to prevent feature interference.

Methods:

Although most medical data sets are not considered IID, it was not possible to tell if this was the case in this particular data set. The patient IDs that were included as one of the features did not refer to the patient, but rather the case. Therefore, I proceeded to split the data under the assumption that the data was IID.

In terms of splitting the data, I generated a function that took the feature matrix, target variable, preprocessor, algorithm, and parameter grid to train and test four different machine learning algorithms. A basic train/test split was performed where 20% of the data was allocated to testing, and the remaining 80% was split using a 4-fold stratified k-fold function. Splits were stratified along the target variable, so that there would be an equal proportion of patients who did and did not experience postoperative wound infection in each train and validation group. Data was also shuffled to mitigate any ordering bias that the data set may or may not have inherently contained. Three different random states were used for each algorithm, and a range of classifier-specific hyperparameters tuned in each random state to output the best model, as determined by the model's accuracy score. Accuracy was used as an evaluation metric because the random undersampling performed during EDA/preprocessing balanced out the dataset. Additionally, through doing an extensive literature review of machine learning in the medical industry, I found that many projects used the accuracy score as their evaluation metric. The chart below details the four machine learning algorithms that were used and their corresponding hyperparameter values.

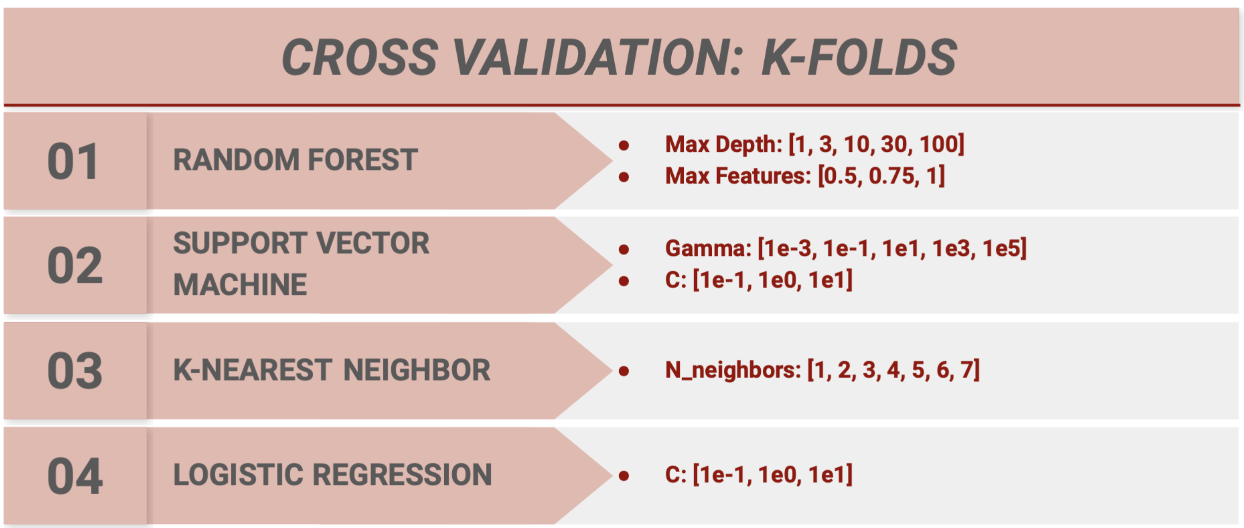

Image 3: The chart above organizes the four different algorithms that data was trained on and their corresponding hyperparameter values.

Results:

Because of the fact that I undersampled my original dataset to balance out the target variable classes and performed a stratified k-fold split along the target variable, the baseline accuracy score was automatically 50%.

In the table below, you can see the performance of each of the models:

Model Performance As Measured by Accuracy Score			
	Mean	*Standard Deviation*	*Standard Deviations above Baseline (0.5)*
Random Forest	0.8111	0.00668	46.57
SVC	0.8081	0.00505	61.01
KNN	0.7626	0.00957	27.44
Logistic Regression	0.7967	0.00538	55.14

All of the accuracy scores were relatively high and close to one another. However, the random forest classifier ultimately outperformed all of the other models, as determined by accuracy score. Since the random forest model had the highest mean and only a marginally higher standard deviation when compared to the next best performing model, support vector classifier, I decided to perform feature analysis on this model.

I first looked at the shap scores of each of the features, and graphed the top scores as shown below.

Image 4: The chart above graphs the top 10 highest shap scores for the random forest model.

What I immediately noticed was that the top two important features were the procedure category and the age that the patient was. This aligns with the exploratory data analysis that I conducted before training any models where I noted differences in the distribution of patients who did not have postoperative wound infection when stratified by age and by procedure category. This observation was further corroborated through calculating the Gini scores of each of the features in the random forest model. In the chart below, you can see that age and procedure category once again rank highly as predictors of postoperative wound infection.

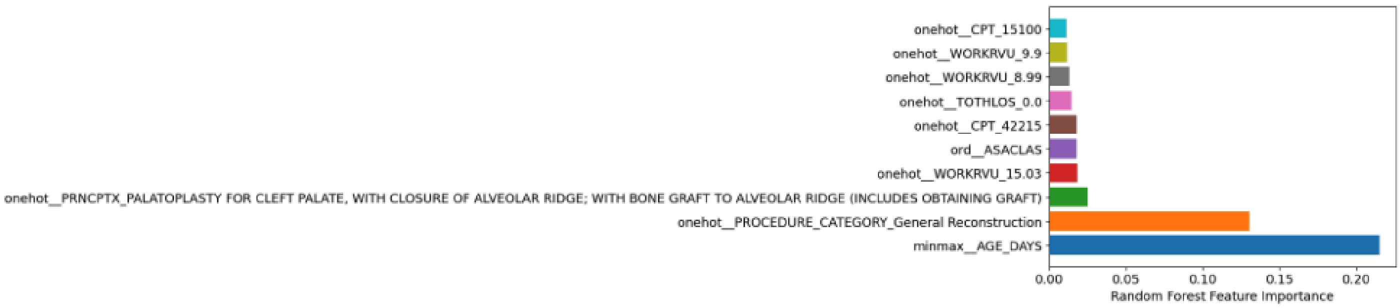

Image 5: The chart above shows the second global feature importance analysis of Random Forest Gini scores.

However, when calculating permutation importance as shown in the chart below, none of the features in this metric align with those found when calculating shap or Gini scores.

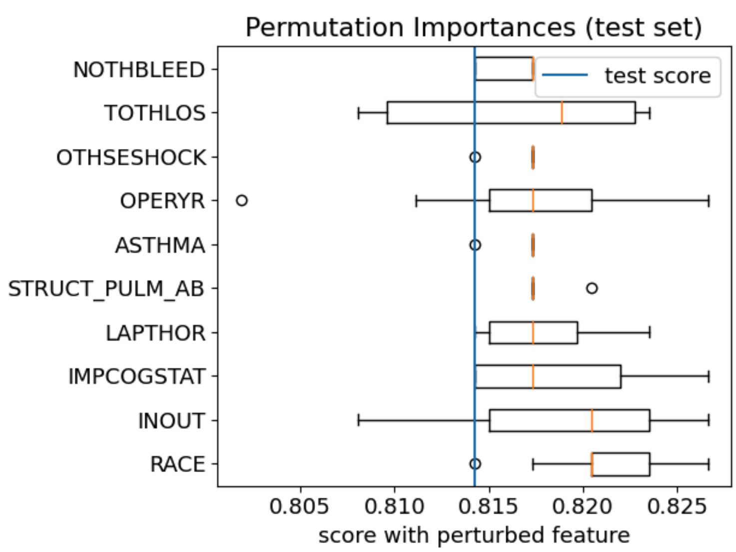

Image 6: The chart above graphs the top 10 most important features as calculated by the permutation importance test.

When force plots were calculated for two randomly selected points, one Class 1 and one Class 0, you can once again see the degree to which age and procedure category played in classifying each point.

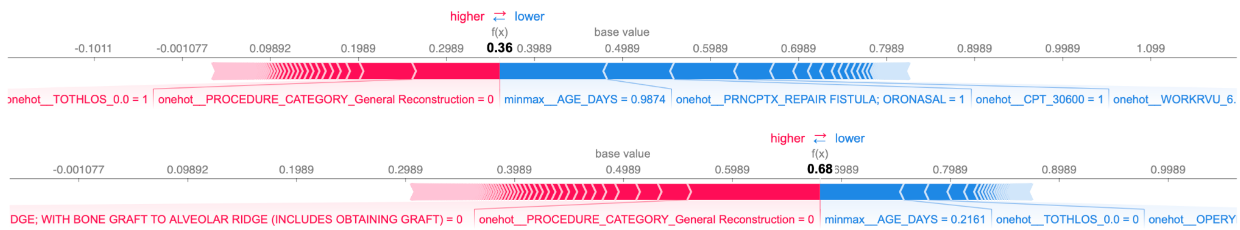

Image 7: The force plots above show the effect of each of the most important features on classifying each point. As you can see, both procedure category and age are closest to the center divide and thus have larger impacts relative to other features when it comes to categorizing features.

From the feature analysis conducted, it appears that whether or not a patient received a general reconstruction category was pivotal in predicting whether or not they would develop postoperative wound infections 30 days later. Additionally, the age of the patient seemed to play a substantial role in the ultimate classification of unknown data. However, the fact that there were no shared features between the shap/Gini calculations and the feature permutation importance graphs is an intriguing revelation that warrants further investigation.

Outlook:

While the results obtained by the four models trained in this particular project are in alignment with other machine learning prediction models in the medical industry, there is certainly room to further improve model predictability and accuracy. Though computationally intensive, it would have been preferable to use the entire data set of over 50,000 patients when training and testing each of the four models. More data could potentially reveal relationships between features and the target variable that were not present in the condensed data set due to the fact that over 80% of the data was excluded from the start.

Additionally, given the condensed nature of the dataset, additional folds and random states could have been used to potentially more accurately report the model's accuracy score. In future iterations, the number of random states should be increased from 3 to 10, as the smaller dataset renders training computationally cheap in this particular situation. However, the low standard deviations and high number of standard deviations above baseline accuracy lead me to believe that despite not having more random states, the models obtained ultimately do a solid job of predicting whether or not pediatric patients will develop postoperative wound infections following surgery. Ultimately, these models are meant more to serve as a helpful predictive tool for plastic surgeons to use in order to minimize suboptimal outcomes for patients, in which accurate classifications are welcome, but not entirely necessary.

References:

Starnoni M, Pinelli M, De Santis G. Surgical Wound Infections in Plastic Surgery: Simplified, Practical, and Standardized Selection of High-risk Patients. Plast Reconstr Surg Glob Open. 2019 Apr 23;7(4):e2202. doi: 10.1097/GOX.0000000000002202. PMID: 31321189; PMCID: PMC6554174.

GitHub:

https://github.com/19sawickin/data1030_final.git

The data file is too large, so github won't allow me to upload it. I even tried zipping the file, and it's still not letting me upload. Additionally, GitHub is messing up the file formatting and not uploading the .ipynb file in the src folder that it's in