PATIENT SAFETY TECHNOLOGY

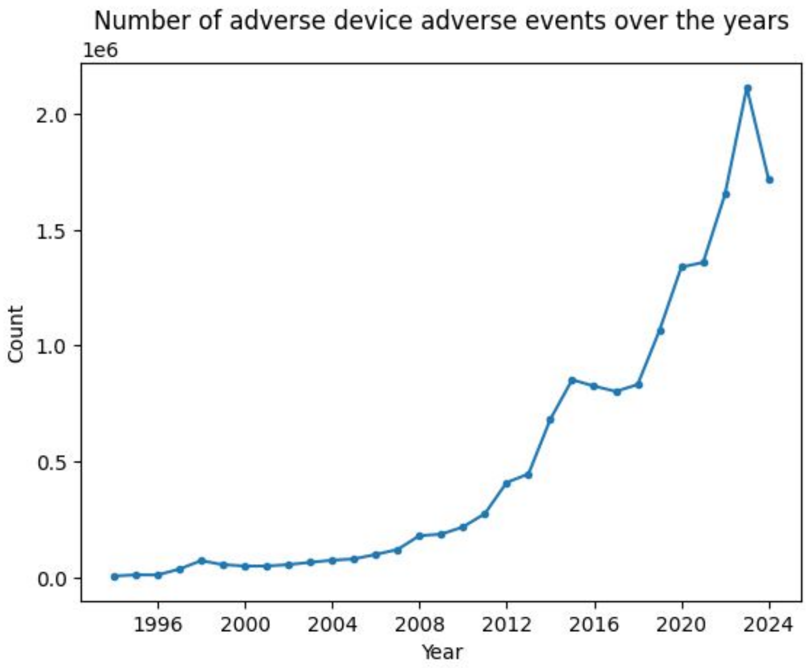
MediSafe: Analysing Safety of Medical Devices & Procedures

9th Annual Data Science Hackathon

Presenters:
Saili Myana, Sai Rithvik Kanakamedala, Rohan Poddar

Problem statement

- There is a persistent problem of medical devices leading to adverse events.
- The objective of this project is to identify patterns in procedure/device-related errors in adverse events associated with medical device usage.
- Number of adverse events involving medical devices is sharply increasing in the last 5 years, and it spiked after the COVID pandemic.
- Unsafe surgical procedures can cause complications in as many as 25% of patients

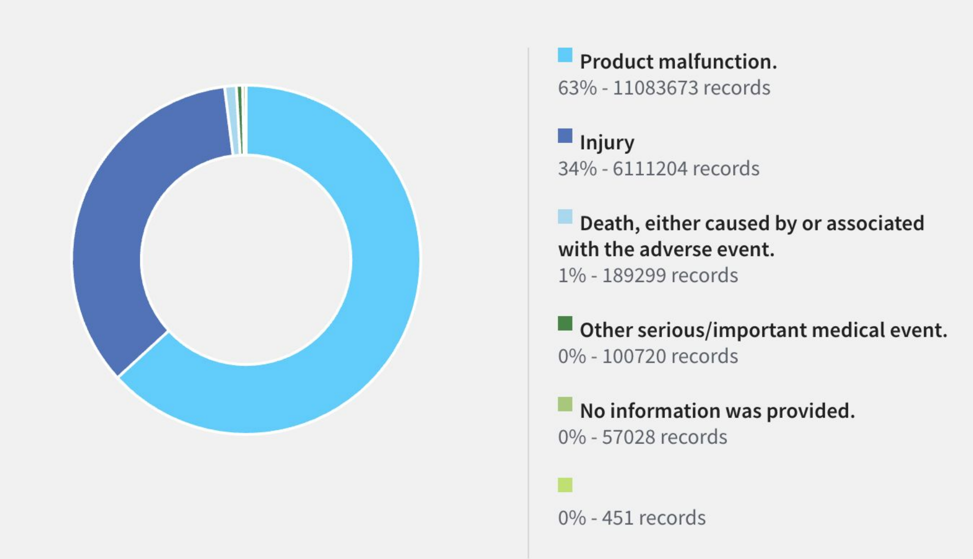


Goal & Novelty

Goal

Through the discussion, we saw that a lot of people from different backgrounds are volunteering and advocating for patient safety. We built a model using the adverse events data so that people can easily input different scenarios,and make inferences from the model.

We want to be aware of the risks of medical devices.



Goal & Novelty

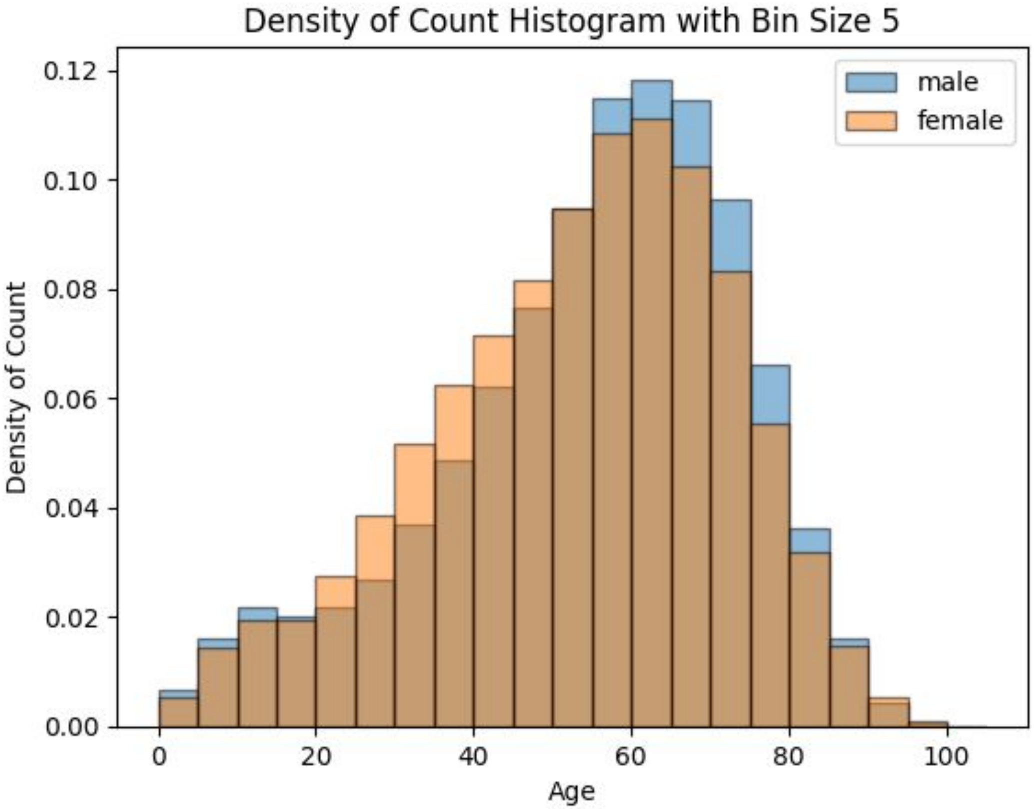
Through this analysis, we want to be able to identify systemic problems in the medical devices.

- **Visualization**
  Analyse various patient and device attributes and their relation to adverse events.
- **Classification & prediction**
  We aim to predict the adverse_event_flag based on the patient and medical device attributes.
  We hope that one day, the adverse events due to medical devices can be successfully predicted by inputting patient demographics

Novelty

- We implemented a classifier that can predict the probability of adverse event.
- Overall, the novelty of our project lies in the combination between visualization and the inferences we can make about mitigating adverse events.

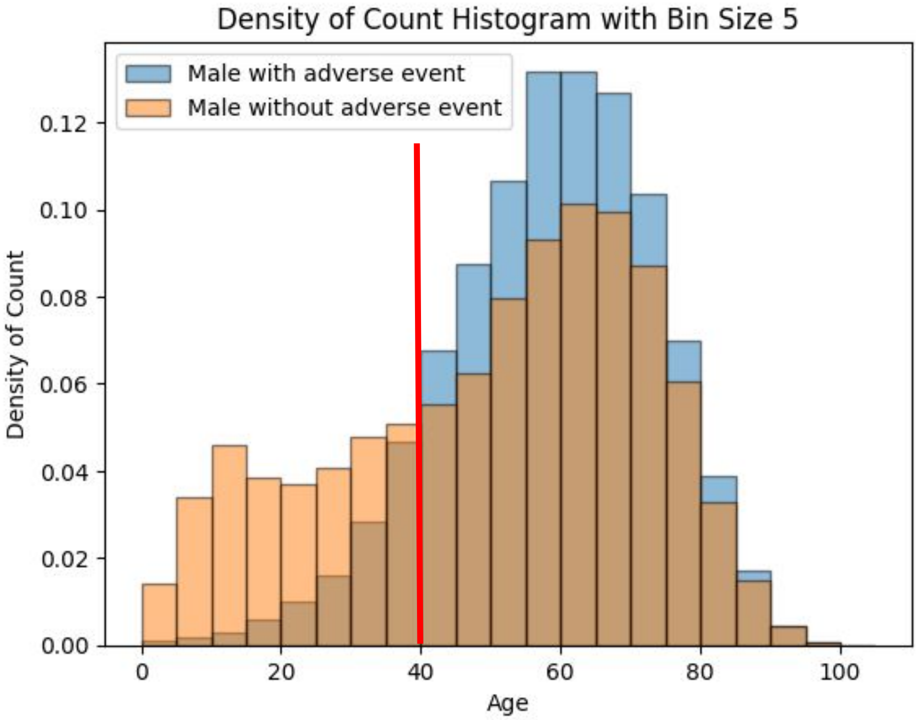
Visualization

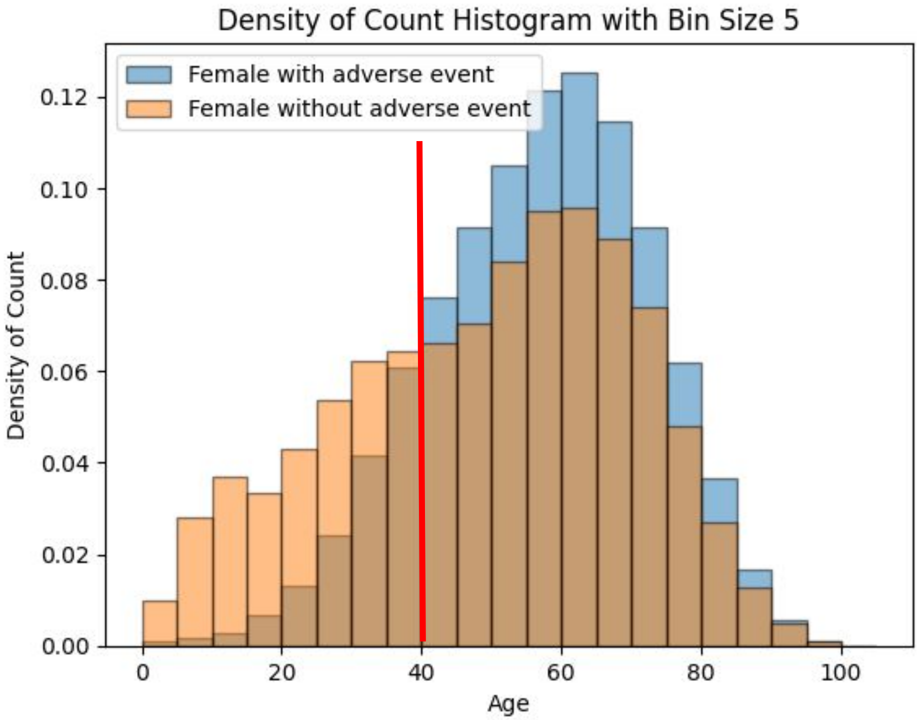


- The density of population for both male and female peaks at 65 years of age.
- Genders: The male and female density looks very similar for all age groups.

* For visualization purposes, we used the entire data for adverse events using API calls. For modelling we only selected a subset of the data to train an initial POC model.

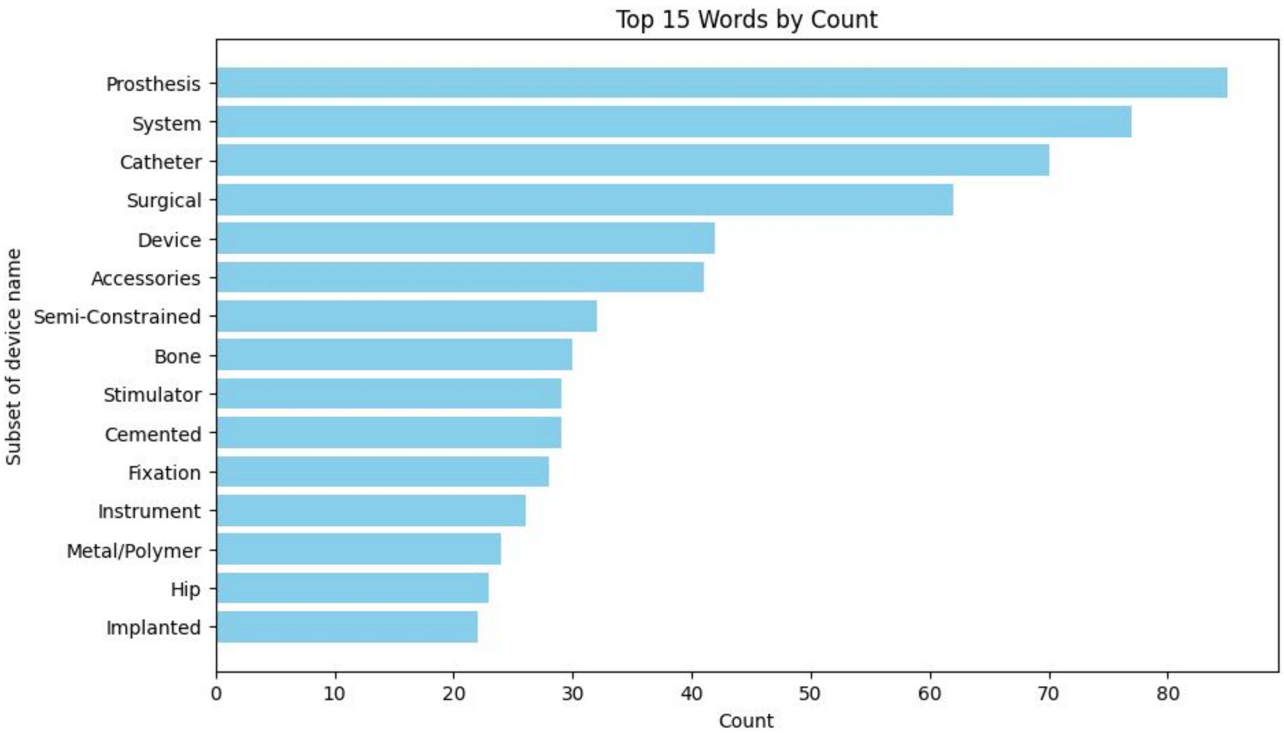
Visualization





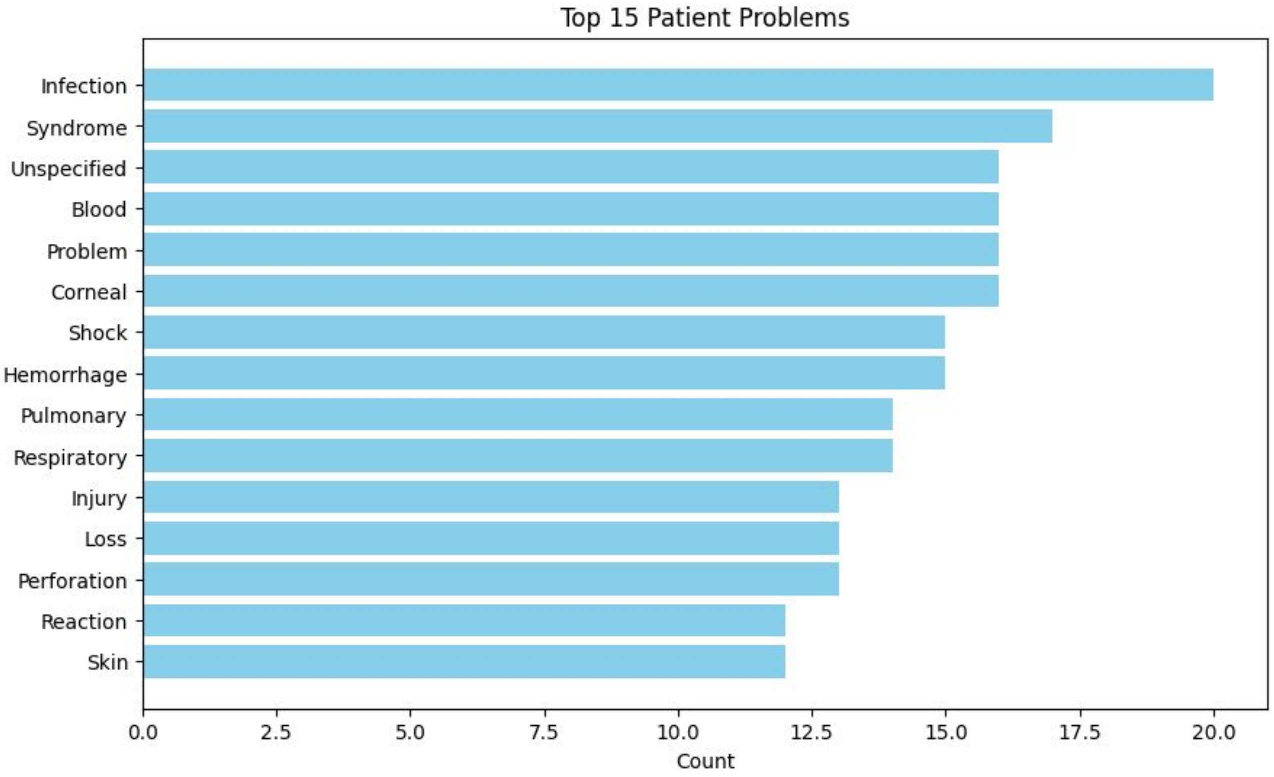
The density of both genders is higher for people with adverse event compared to people without adverse event. We can imagine a boundary at age 40 before which density is higher for male without adverse event.

Visualization



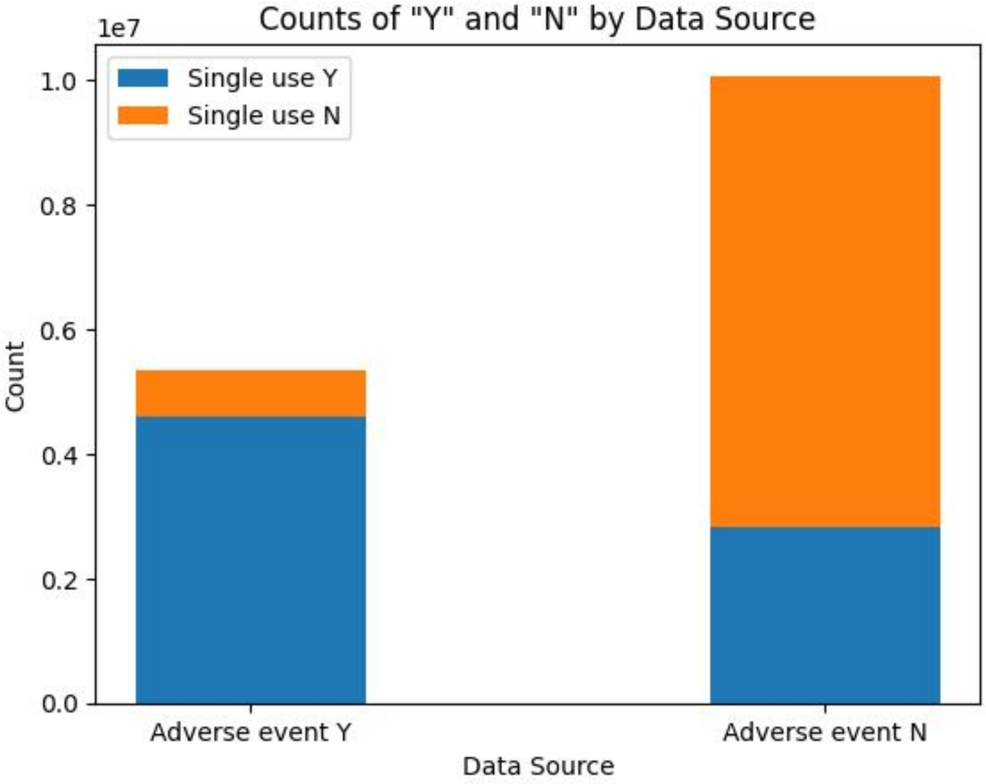
- When **adverse event flag = ‘Y’** most common medical devices are plotted.
- Prosthesis, Catheter, surgical device, are among the top occurrences when adverse event flag = Y

Visualization



- When **adverse event flag = ‘Y’** most common medical devices are plotted.
- We can see that Infection, blood problem, corneal problem, hemorrhage are among the top patient problems when adverse event flag = Y

Visualization



- We can observe that among the reports with adverse event = Y count of single use = Y is higher than single use = N

- Similarly, among adverse event = N count of single use = N is higher than single use = Y

Dataset preprocessing

Data for Classification :

	reprocessed_and_reused_flag	previous_use_code	product_problems	health_professional	single_use_flag	device_operator	device_evaluated_by_manufacturer	openfda.medical_specialty_description	openfda.device_class	sequence_number_outcome	patient_sex	patient_problems
0	N	R	[Pumping Stopped]	N	N	Unknown	Y	Clinical Chemistry	2	UNK	No Answer Provided	[No Clinical Signs, Symptoms or Conditions]
1	UNK	I	[Insufficient Information]	UNK	Y	LAY USER/PATIENT	R	General Hospital	2	UNK	Female	[Purulent Discharge, Hyperglycemia, Pain, Skin...
2	N	I	[Migration]	Y	Y	HEALTH PROFESSIONAL	R	Unknown	3	Other	Male	[Inadequate Pain Relief]

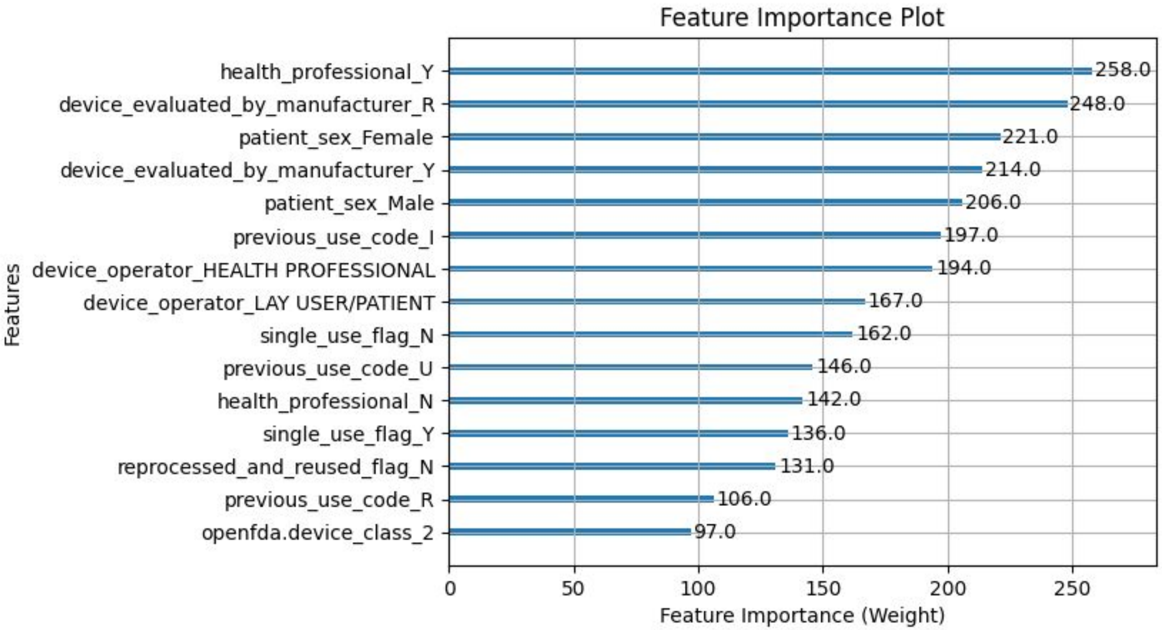
- **Data**
    - The data we trained on contains 10035 rows, each corresponding to an adverse event.
    - We considered the 10 categorical columns shown above for the classification task.
    - These columns were one-hot encoded and patient_problems and product_problems were dropped. We hope to use language models to embed these features later on.
    - We used a subset of the data (first part of first quarter of 2023) for our current analysis, we hope to expand it in the future.

Why XG Boost?

- Implemented an **XGBoost** classifier model that learns incrementally.
- We also compared a few other models and found XGBoost to be the best suited for our use-case.
- **Scalability**: suitable for large datasets, hence best suited for our use-case.
- **Customizability**: wide range of hyperparameters that can be adjusted.
- **Missing values**: can handle missing values, our dataset has lots of missing fields.
- We achieved a validation accuracy of **92.7%** with the help of the XGBoost classifier. Though the target feature “adverse_event_flag” is slightly imbalance with approximately 66%Y and 24%N, the classification report below shows the model is able to generalize well on the test set.

	precision	recall	f1-score	support
0	0.94	0.94	0.94	12481
1	0.90	0.90	0.90	7526
accuracy			0.93	20007
macro avg	0.92	0.92	0.92	20007
weighted avg	0.93	0.93	0.93	20007

Results



- Health care professional feature was found to be the most important to predict adverse event. It is expected that when medical devices are used in the presence of health professionals, the rate of adverse events is lower.
- Similarly, we can infer from the plot that whether or not the device was operated by the patient or professional impacts the prediction.
- It is also interesting to note that patient sex is also an important feature for prediction.
- Device class, a classification of risk/ seriousness of the device is also one of the important predictors.

Future Work

- Ideally, we want to be aware of all the risks in the medical devices that are used.
- We also want to be able to predict any adverse events for specific patients based on their demographics and health history.
- We can further expand the classification task on the basis of type of adverse events and perform a multi class classification.
- The dataset also has many text columns and we want to embed them to capture the information into embedding in the future.
- The dataset we used corresponds to one of the json files.The remaining json files are saved in google drive. Our notebook can also be used to clean all the other json files. The clean csv files can now be combined to get the entire dataset. This way, our model also is scalable.

References

1. https://www.patientsafetytech.com/
2. https://open.fda.gov/

THANK YOU :)