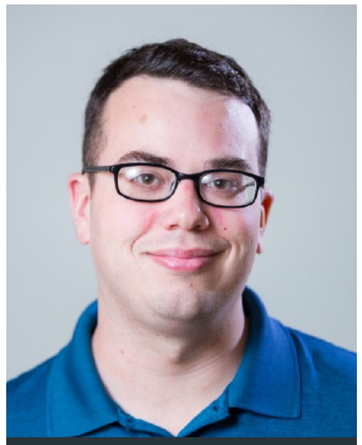
ICU SURVIVED

PREDICTING ICU SURVIVAL

CURIECAPSTONE.COM

THE DATA SCIENCE TEAM THE DOCTOR ORDERED



Shay Altshue

in/shay-altshue
/shayaltshue
shayaltshue@gmail.com

Ravinder Singh

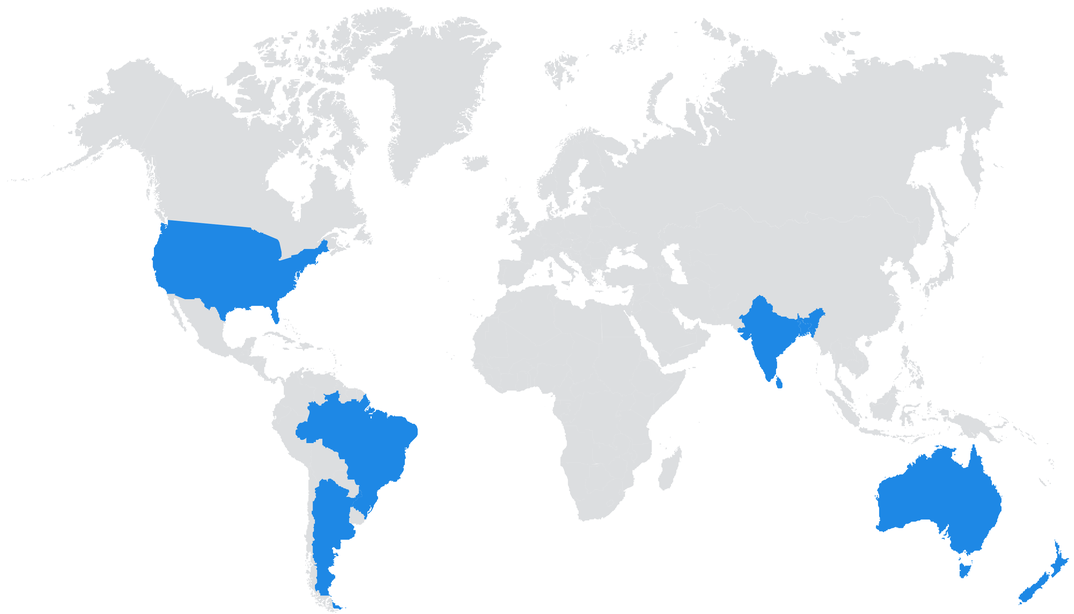
in/ravinder-d-singh
/ravinder79
ravinder79@gmail.com

Yipeng Jiang

in/yipeng-jiang
/jiangyipeng718
yipengjiang0532@gmail.com

Chase Thompson

in/wchasethompson
/chasethompson
chase.w.thompson210@gmail.com



Our team created a machine learning model to predict and identify trends in patient survival following ICU care.

We used data provided by MIT-led consortium (GOSSIS) from 91,000+ patients, from 200 hospitals in 9 countries.

Data Science Tools Used

Programming Language
Python

Visualization
Matplotlib, Seaborn, Plotly, shap

Frameworks & Libraries
Pandas, Numpy, Scikit-Learn, SciPy Stats, XGBoost, LightGBM

Other tools
Jupyter Notebook, Github

8
COUNTRIES

200
HOSPITALS

91K
PATIENTS

185
VITALS

DATA SCIENCE PIPELINE & RESULTS

PLANNING & DELIVERY

Create a classification model which predicts ICU patient survival. Our goal was to create a model which is better than the current model (aka APACHE model) used by the hospitals around the world today.

DATA WRANGLING

91,000 + patient data. For each patient, there are 185 columns (independent variables)

CHALLENGE

30% of data is missing

OUR STRATEGY

- Statistical Imputation

- Sklearn iterative imputation

- Models which natively handle null values (XGBoost, LightGBM)

- Domain knowledge to fill the missing values

EXPLORATION & FEATURE ENGINEERING

Form initial hypothesis on predictors and do hypothesis testing e.g. Chi2 test, Pearsons' correlation test

CHALLENGE

Create new features which are better predictors than obvious ones and an imbalanced dataset

OUR STRATEGY

- *Consult with domain experts (M.D, ICU nurse/ techs) and created new features which provides better outcome separation*
- *Choose appropriate evaluation metric (ROC- AUC)*
- *Tried SMOTE or the Synthetic Minority Over-sampling Technique.*
- *Tried Anomaly detection approach*
- *Choose tree-based model*

MODELING

Created different classification models:

- Logistic Regression

- Decision Tree

- XGBoost

- LightGBM

Our best model (LightGBM) preformed 7% better than the baseline on out-of-sample data. 7% better translates to 1,623 fewer incorrect predictions in patient population of 90k

CHALLENGE

Avoid model overfitting

OUR STRATEGY

- Used K-fold cross-validation to determine best hyper-parameters without overfitting

RESULTS

PER PATIENT RISK CLASSIFICATION

Our model allows physician to visualize underlying risk factors for patients on individual basis

TWO IMPORTANT PATIENT GROUPS

Identify two key patient cohorts: high risk patients and unexpected outcome patients is missing

OUR EXPLAINABLE MODEL

Our Tree-based models have explainable functionality

DEEP DIVE

Find the deck, interactive charts, a deeper dive explanation of the data and model, and more