Predictive Analytics Capstone Project

Presented by: Lee Ping Tay

Introduction

Project Goal

Predict will a critically ill patient who undergoes Right Heart Catheterization (RHC) die during the procedure.

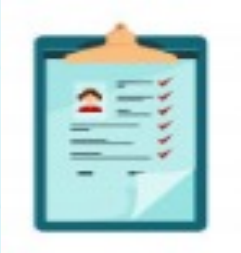
Right Heart Catheterization (RHC)

In a RHC procedure, the doctor passes a thin tube into the right side of the patient's heart to check conditions of heart and lungs.

Dataset

Source of the data

Department of Biostatistics at Vanderbilt University.



Number of records

About 5700 hospital admissions from critically ill patients.
About 2100 of them received a RHC procedure.

Contents

Patient's personal and health information.
e.g., age, disease category, and admission diagnosis.

Prevalence

Definition

Proportion of population that is positive.

Prevalence for the dataset =

$$\frac{\textit{Total number of mortality rate}}{\textit{Total population of critically ill patients who received a RHC}}$$

Prevalence measured from the data = 0.68

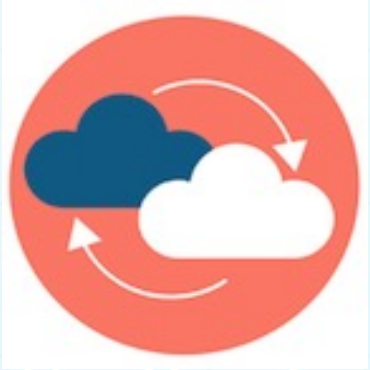
68 percent of the critically ill patients who undergo Right Heart Catheterization (RHC) will die during the procedure.

Steps for Classification Model

Select data

Train model

Validate model



Tune model

Test model

Measure performance and make prediction

Features Engineering

Numerical

34 numerical variables about measurements of patients' health and illness (e.g., blood pressure, heart rate, Glasgow coma score)

Categorical

Five categorical variables about patients' primary and secondary disease category, and cancer.

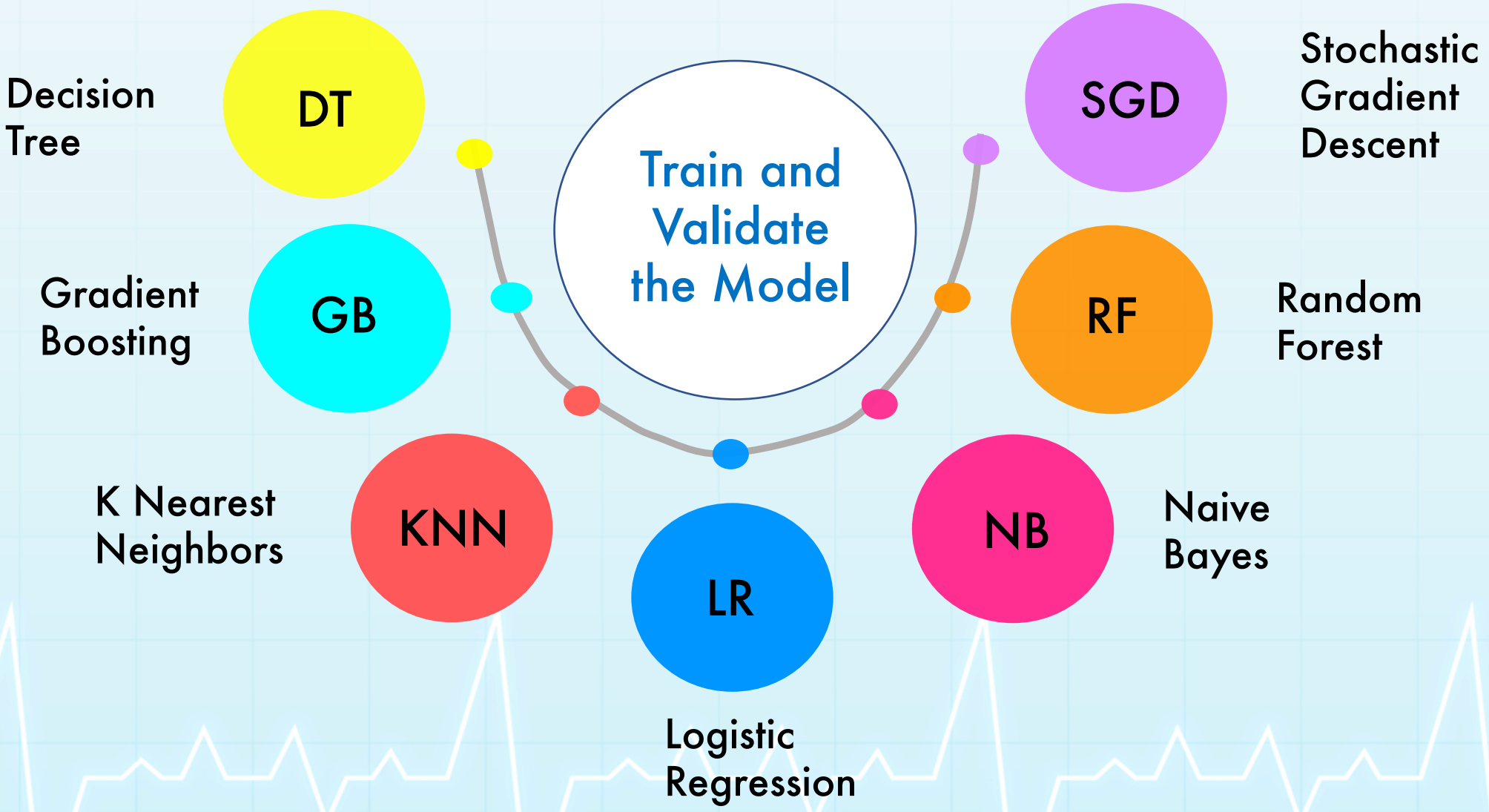
One-hot encoding to convert categorical data to numerical data.

Ordinal

Two ordinal variables about age and income.

Convert ordinal variables to numerical variables by using ranges and assigning an id.

Model Selection

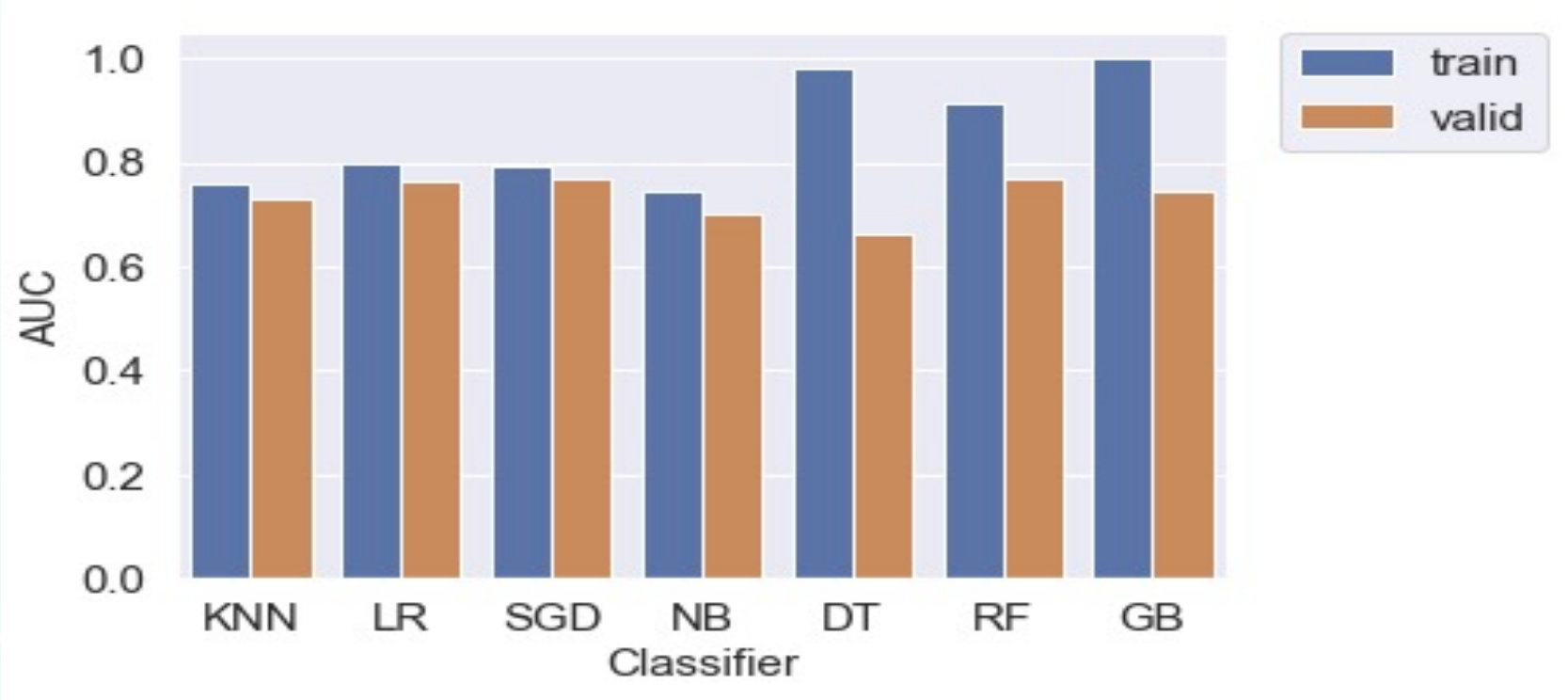


Baseline Model Performance

Best Baseline Model : Stochastic Gradient Descent (SGD)

Summary

- SGD has the highest AUC score of 0.77 in validation dataset.
- SGD also has less overfitting.

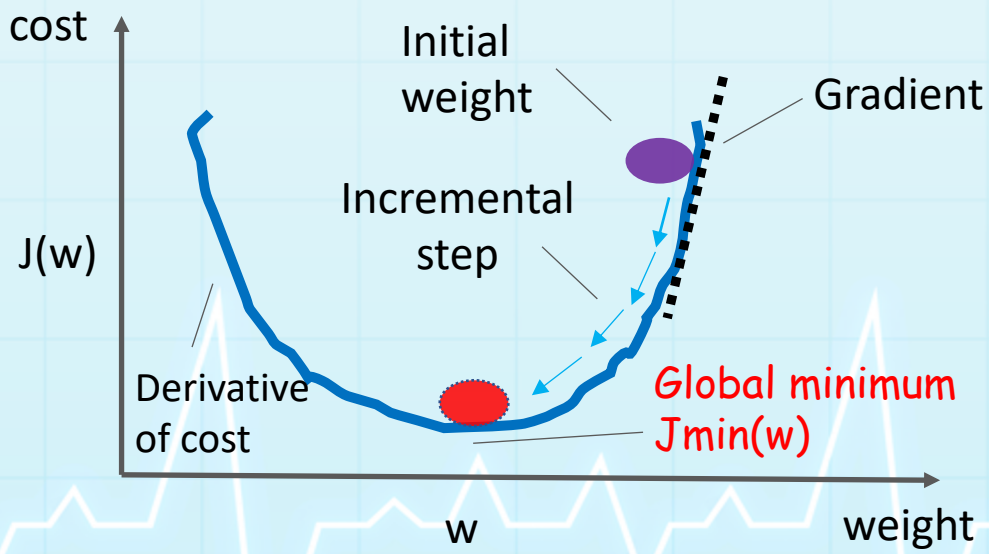


Stochastic Gradient Descent (SGD) Model Algorithm

SGD is faster than Logistic Regression (LR)

SGD calculates the gradient (minimum of a function) for each step by using a random sample data point at one time then moves forward and repeat the steps.

LR with gradient descent looks at all the training data points in every iteration.



descent algorithm visualized (Lanham 2018)

Hyperparameter Tuning

Parameter for SGD Model

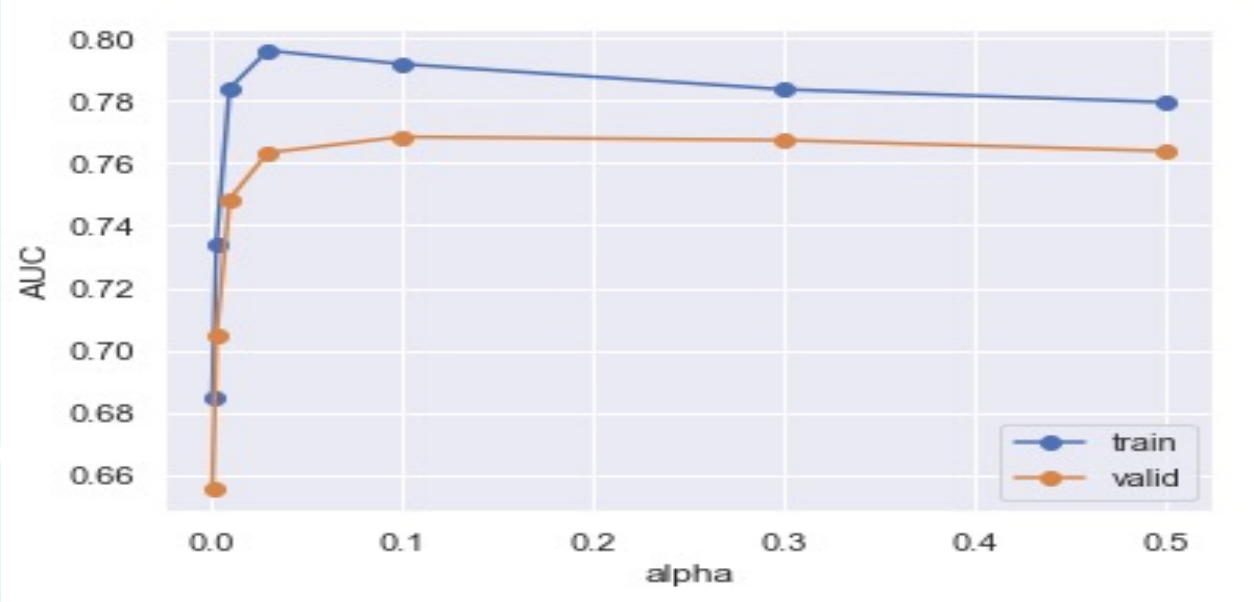
Alpha (also known as learning rate or step size)

It is the portion that weights are updated when we train the model.

Effect of Alpha for SGD Model

Range of alpha [0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 0.5]

As we increase the alpha from 0.001 to 0.03, we get improved AUC in both the training and validation performance. However, the performance decreases after the alpha of 0.1.



Techniques for Improving Models

Learning Curves

Reduce overfitting

(e.g., decrease number of input features)

Feature Importance

Logistic Regression
Top Positive and Negative feature

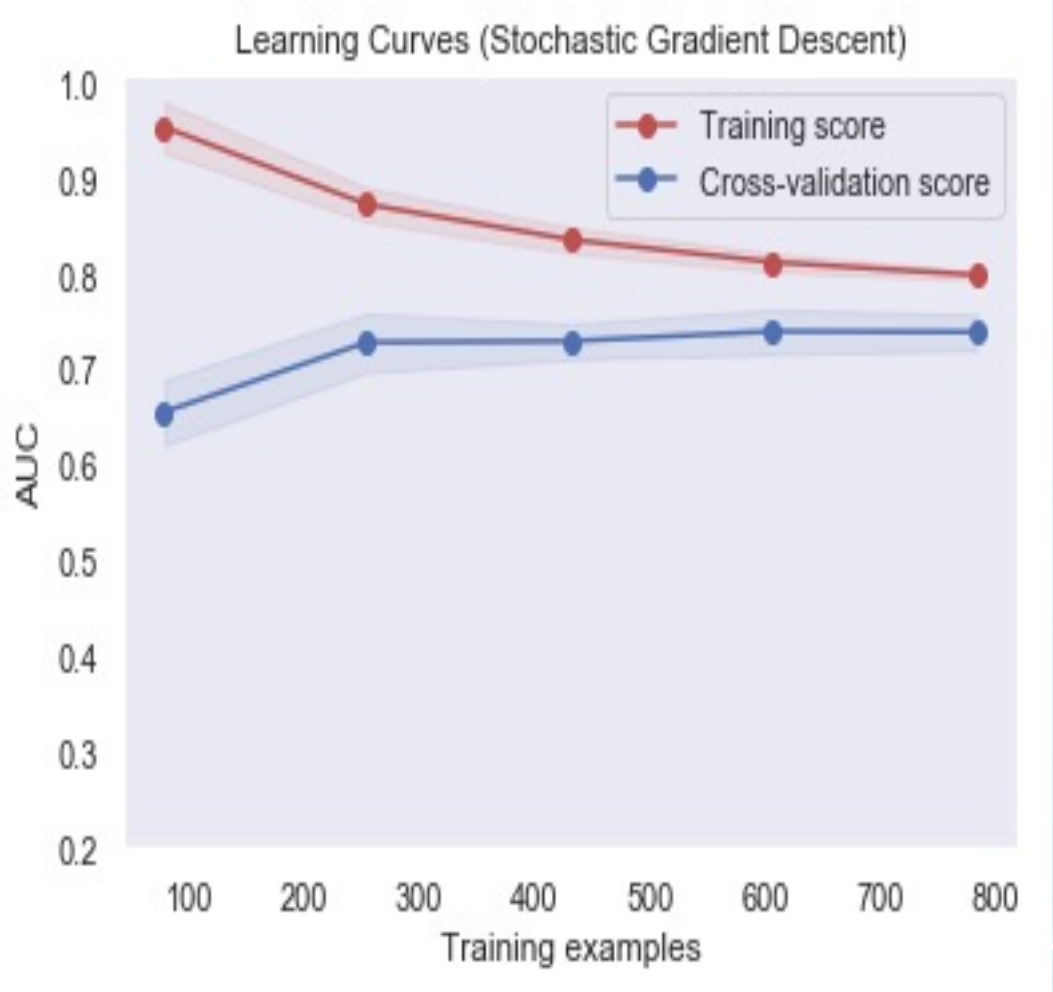
Random Forest
Top Positive feature

Hyperparameter Tuning

Randomized Search Cross Validation

Find best hyperparameters

Learning Curves for SGD Model



Problem

high variance (overfitting) as the learning curves converge to a low score.

Possible Cause

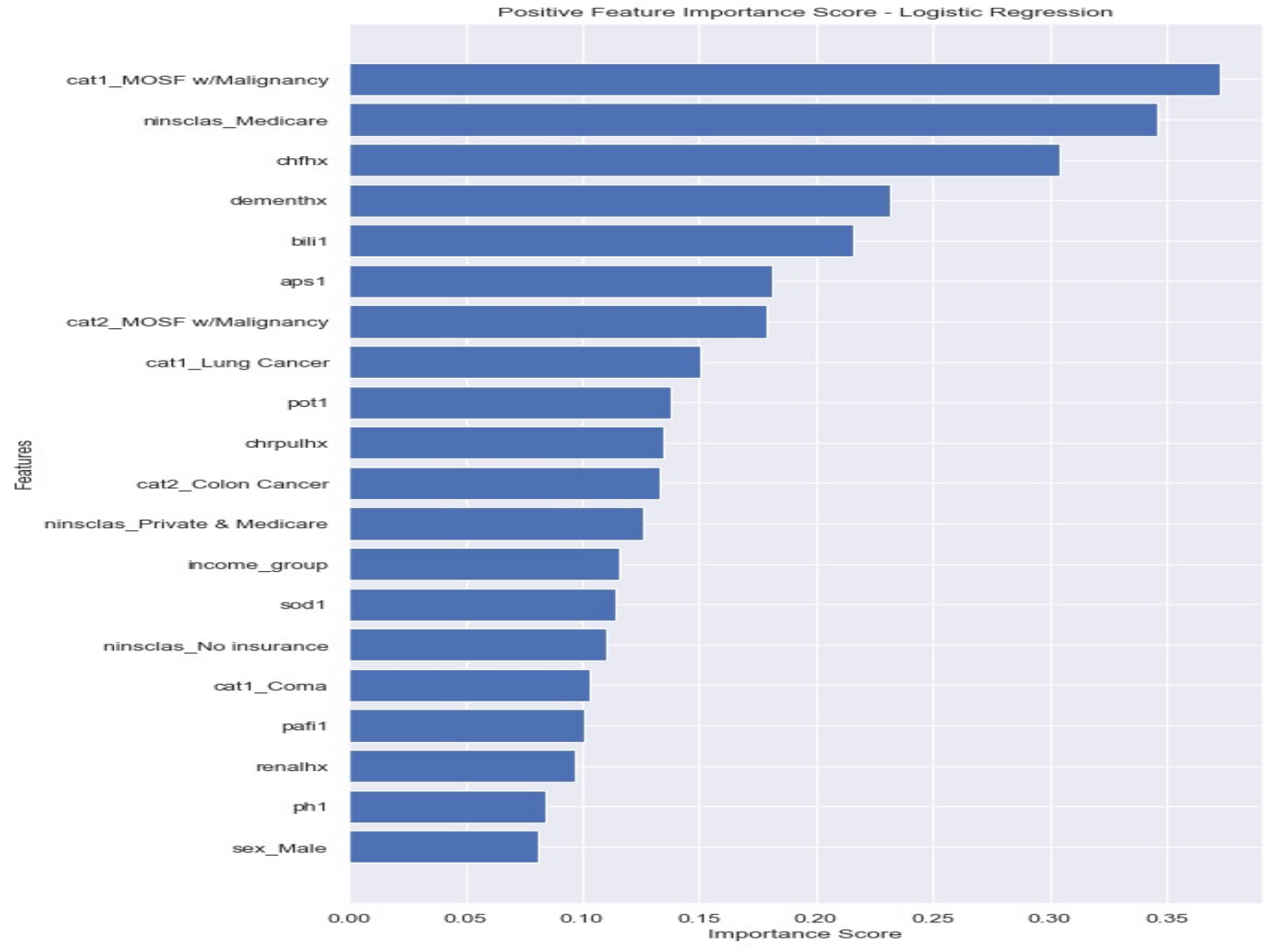
Model is too complex

To reduce overfitting

- Decrease number of input features
- Increase regularization
- Introduce better features
- Simplify the model

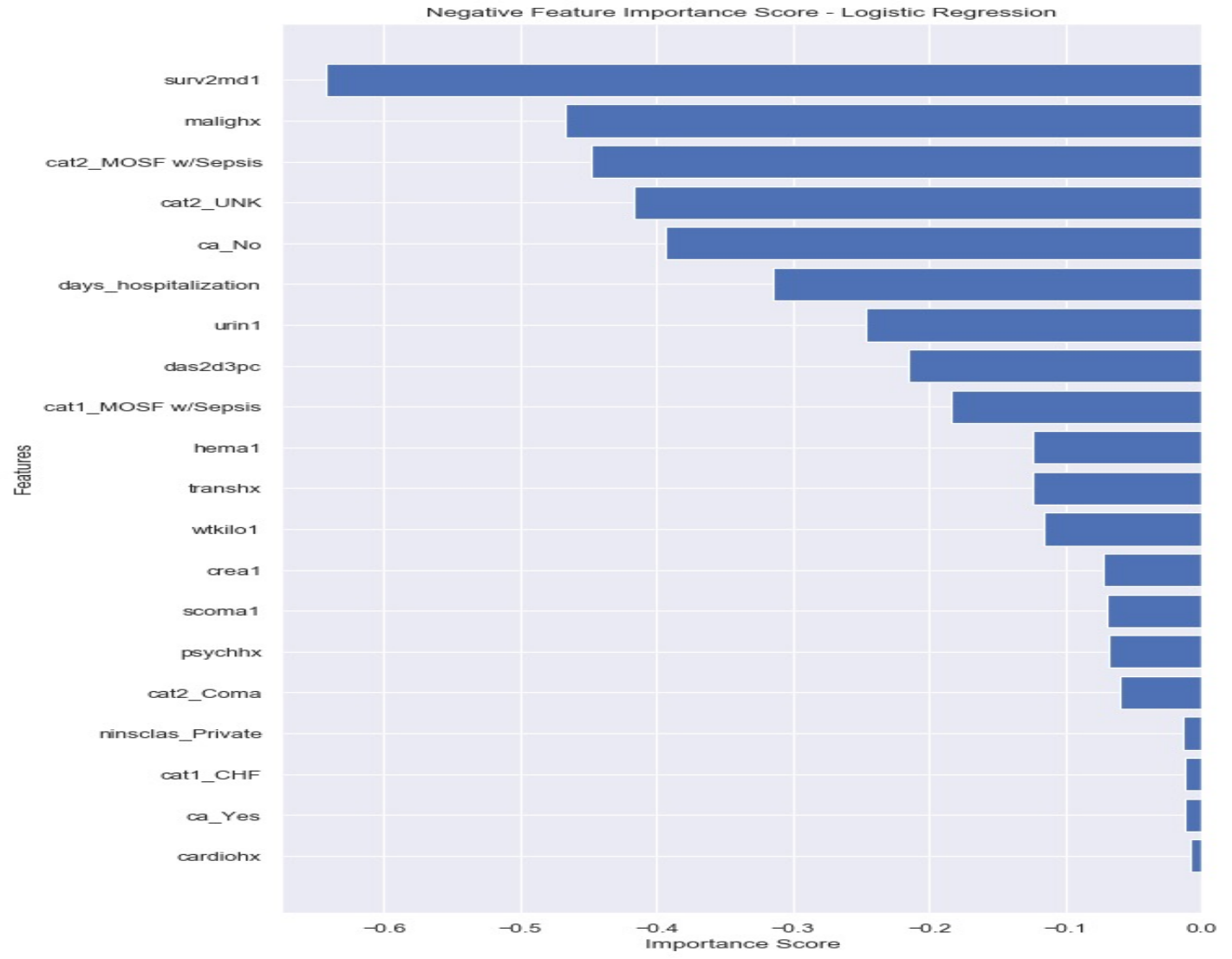
Top Positive Feature Importance for Logistic Regression

The most important positive feature is multi-organ system failure (MOSF) with malignancy (cancerous).



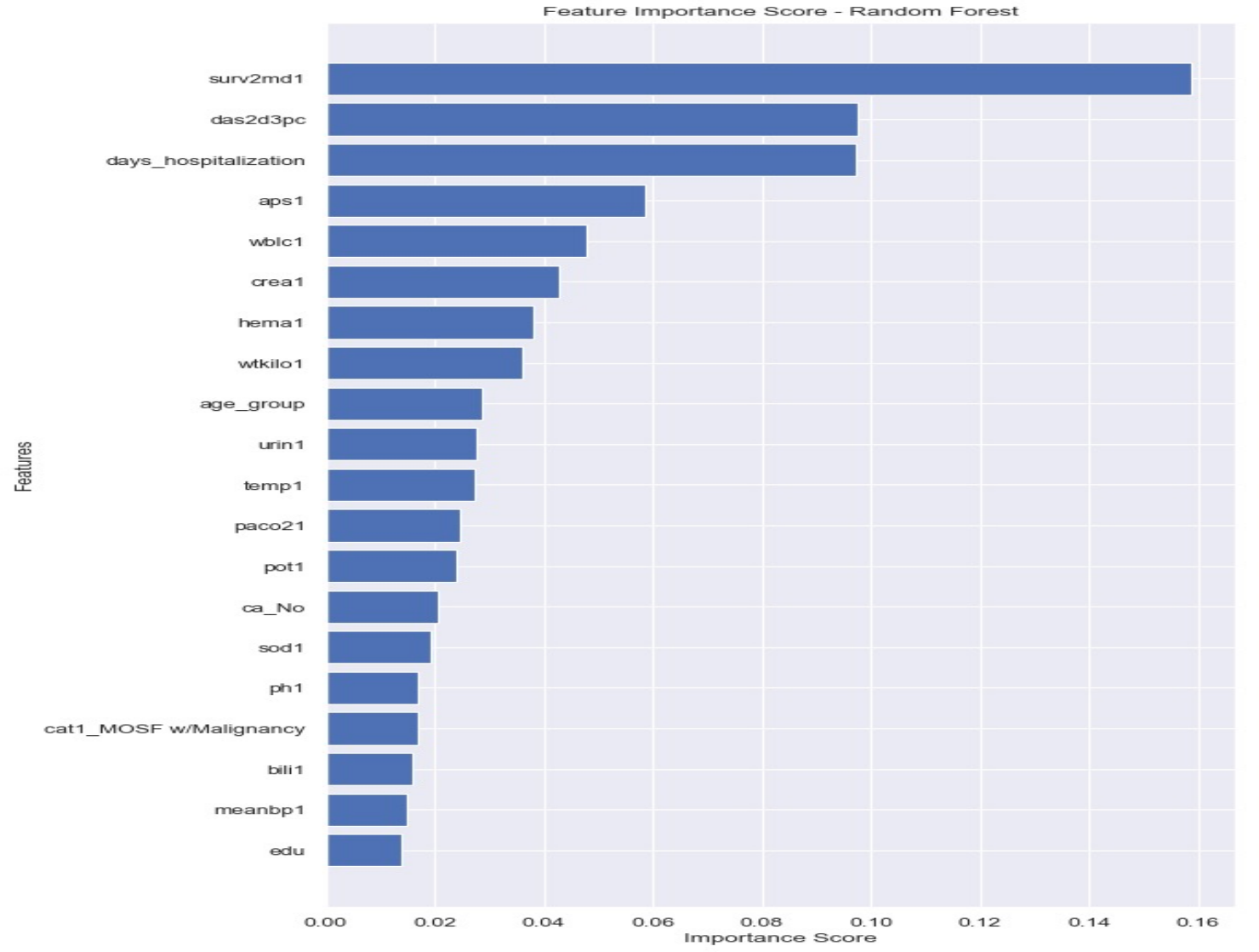
Top Negative Feature Importance for Logistic Regression

The least important negative feature is estimate of the probability of surviving 2 months at day 1.



Top Positive Feature Importance for Random Forest

The most important positive feature is estimate of the probability of surviving 2 months at day 1.



Hyperparameter Tuning

Use Randomized Search Cross Validation to optimize four baseline models
Stochastic Gradient Descent (SGD), Random Forest (RF), Gradient Boasting (GB), and Logistic Regression (LR).

Best hyperparameters for Optimized SGD

penalty: 11, max_iter: 800, alpha: 0.03

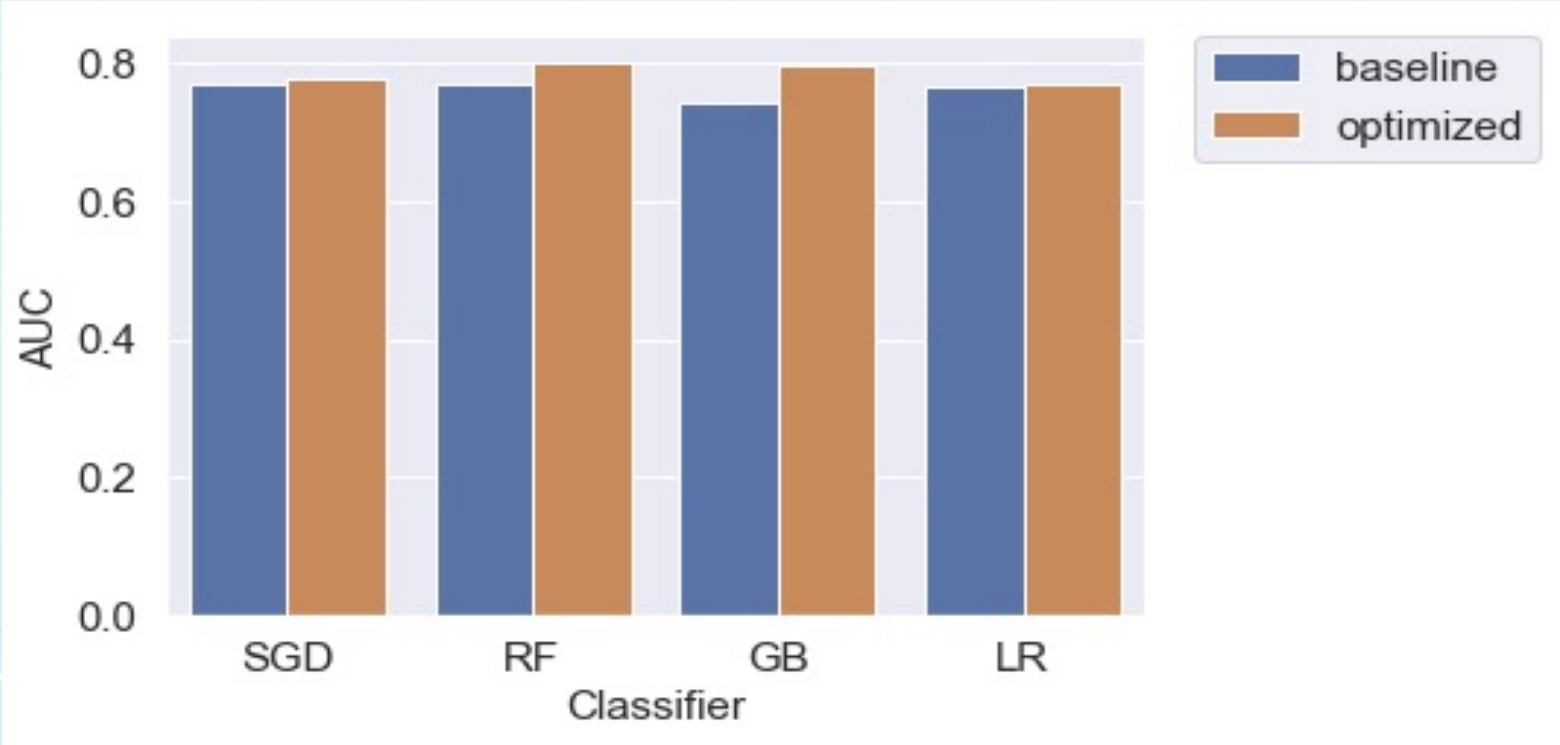
Optimized SGD receives a boost of 0.01 on the Validation AUC.

	Model	Dataset	AUC
1	SGD	baseline	0.77
2	SGD	optimized	0.78
3	RF	Baseline	0.77
4	RF	optimized	0.80
5	GB	baseline	0.74
6	GB	optimized	0.79
7	LF	baseline	0.77
8	LF	optimized	0.77

Model Selection: Analyze Optimization Results

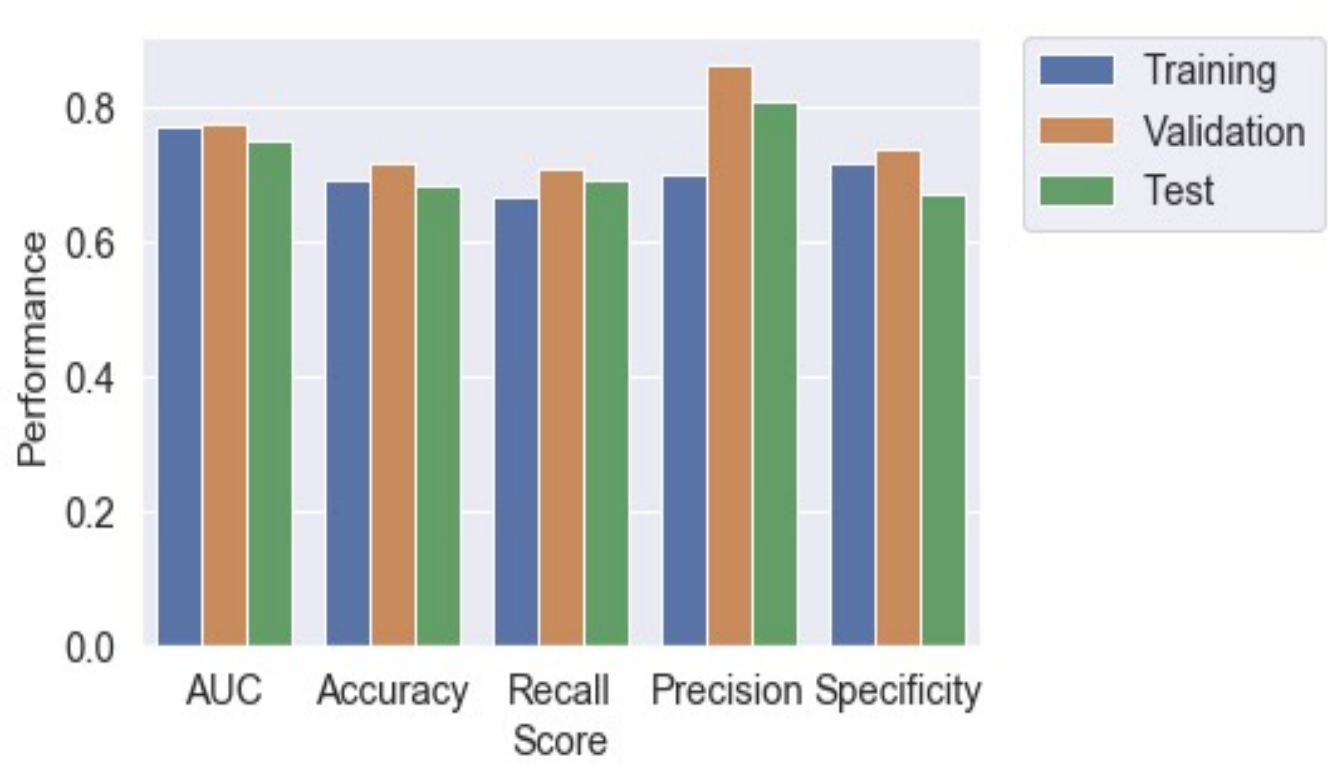
Best model - Optimized Stochastic Gradient Descent (SGD)

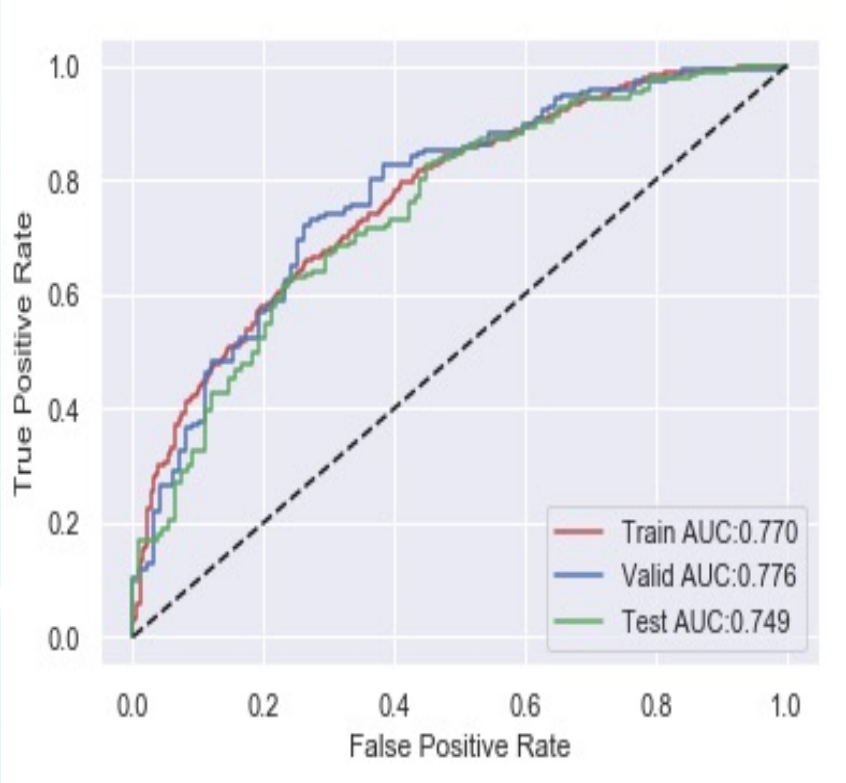
Optimized SGD has the closest Training versus Validation scores and its performance is about the same as the RF, GB, and LR classifier.



Model Evaluation

Based on the Test dataset, optimized SGD has AUC of 0.75 that would catch 69 percent of the death rate when we use a threshold of 0.5.





Conclusion

In this project, we predict will a critically ill patient who undergoes a right heart catheterization (RHC) die during the procedure. We trained seven models and analyzed the performance.

Lift ratio (Precision/ Prevalence) = 1.2 times.

Our optimized SGD model is able to capture 1.2 times more patients compared to picking the same number of patients randomly.

References

Dupont, C. (2018, April 26). Right heart catheterization dataset. Retrieved from http://biostat.mc.vanderbilt.edu/DataSets

Freepik. big-data-analytics-technology-flat-icons-set-with-internet-cloud [Image file]. Retrieved from https://www.freepik.com/free-vector/

Freepik, close-up-stethoscope-cardiogram-chart [Image file]. Retrieved from https://www.freepik.com/free-photos-vectors/

Freepik. heart-shape-with-bandages-stethoscope-ecg-analyzer-report-wooden-table [Image file]. Retrieved from https://www.freepik.com/free-photo/

Freepik. People vector [Image file]. Retrieved from https://www.freepik.com/free-photos-vectors/

Healthcarestocks [Image file]. Retrieved from https://biotechinvestingfinance.files.wordpress.com/

Lanham, M. (2018, March 30). Learn ARCore - Fundamentals of Google ARCore: Learn to build augmented reality apps for Android, Unity, and the web with Google ARCore 1.0. Retrieved from https://www.oreilly.com/

Starline. Background vector [Image file]. Retrieved from https://www.freepik.com/free-photos-vectors/