Prediction of Patient Survival using Supervised Machine Learning

Leena Alafifi, Mamoon Uddin, Kaytlin Simmer
Mentor: Dr. Nouhad Rizk

College of Natural Sciences and Mathematics, University of Houston

Abstract

For our research project we are using supervised machine learning to predict the likelihood of the survival of patients admitted to a hospital. We are using data compiled from patient records for comparative analysis of neural networks and logistic regression to ultimately predict in hospital patient death. Identifying the probability of patient mortality will help doctors recognize particular risk factors for patient death and handle higher risk cases with more intensive care.

Patient mortality rates are commonly used to measure hospital performance and the overall goal for any medical center is to decrease their rate of patient death. Being able to accurately predict the survival of a patient based on their vitals and inputted factors from their records is one of the first steps to be taken to improve patient care and increase patient survival.

Background

Deep-Learning-Based Survival Prediction of Patients in Coronary Care Units study done in 2021 by Rui Yang, Tao Huang, Zichen Wang, Wei Huang, Aozi Feng, Li Li, and Jun Lyu. Methodology: neural networks

The accuracy of the neural network model (0.822) was about 4% better than that of the CPH model (0.782). It was indicated that deep learning may be more suitable for handling large samples, multivariate and nonlinear survival analysis than the CPH model. Based on these studies we have chosen to implement neural networks, logistic regression, and random forests. To go a step further, we will implement hyperparameter tuning to improve prediction accuracy. We are also using a much larger dataset.

The dataset we are using is a public dataset compiled by Mitisha Agarwal. This dataset contains 85 variables including patient vitals and presence of risk factors such as chronic disease. It contains overall 91,713 data entries collected from patient records. For our use: We dropped the 7 columns of ‘apache_3j_bodysystem’, ‘apache_2_bodysystem’, 'Unnamed: 83','encounter_id', 'patient_id’, 'hospital_id', 'icu_id’ and are overall using 57, 598 data entries for our predictive modeling.

Methods

Hyperparameter tuning

Hyperparameter tuning consists of finding a combination of hyperparameters that maximizes the model’s performance, minimizing a predefined loss function to produce better results with fewer errors. For Hyper Parameter tuning we used GridSearchCV to get the best model based on accuracy.

Neural Networks

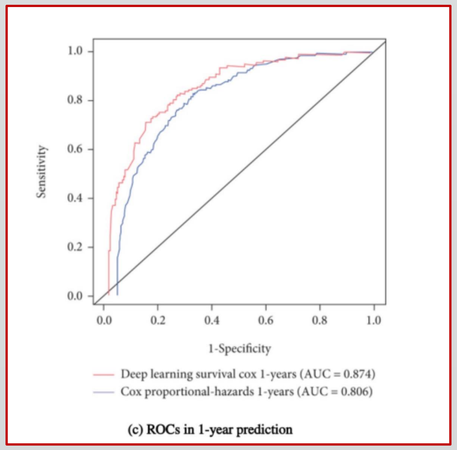
A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain through use of perceptron's. To make the Neural Network we first Split and scaled the data using a standard scalar, then built the Sequential model and fit the model.

Logistic Regression

Logistic regression is a statistical supervised machine learning model used for classification and predictive analysis. To create this model, we split the data into test and train sets, used a standard scaler for more accurate results, and used sklearn to implement test_train_split() to randomize the test and train subsets. We then defined the model and parameters to determine the optimal logistic regression parameter, which was ‘C’: 1.0, ‘penalty’: ‘12’, ‘solver’: ‘ newton-cg’. We then fit and implemented the model using this parameter.

Random Decision Forests

Random forest modeling utilizes multiple decision trees (which build a “most likely” relation between inputs) and takes the average of the outcome received to classify. Due to the large quantity of the dataset, the computation time for random forests was not ideal for our research.



(c) ROCs in 1-year prediction

Results

For the Evaluation of the models we looked at the Classification Report, Confusion matrix, Mean Square/ Absolute Error Accuracy Score, F1 Score, Recall Score , Precision, and AUC-ROC curve.

Logistic Regression

	precision	recall	f1-score	support
0	0.93	0.99	0.96	15787
1	0.63	0.26	0.36	1493
accuracy			0.92	17280
macro avg	0.78	0.62	0.66	17280
weighted avg	0.91	0.92	0.91	17280

Mean Absolute Error: 0.07731481481481481
Mean Square Error: 0.07731481481481481
Accuracy Score: 0.9226851851851852
Recall Score: [0.98574777 0.25586068]
precision Score: [0.93336532 0.62932455]
AUC-ROC Curve: 0.6208042251691992

Neural Network

	precision	recall	f1-score	support
0	0.92	0.99	0.95	15781
1	0.48	0.10	0.17	1499
accuracy			0.91	17280
macro avg	0.70	0.55	0.56	17280
weighted avg	0.88	0.91	0.89	17280

Mean Absolute Error: 0.08755787037037037
Mean Square Error: 0.08755787037037037
Accuracy Score: 0.9124421296296297
Recall Score: [0.98954439 0.10073382]
precision Score: [0.92053761 0.4778481]
AUC-ROC Curve: 0.5451391056851833

We could see that the logistic regression performed better. Even though Logistic regression seems to better. the confusion matrix shows that neural network is better cause the False positive % is way lower, which is more clinically significant in our findings.

Conclusion/Future Direction

We found neural networks to be the most reliable form of supervised machine learning to predict in hospital death.

Inpatient mortality rates in the United States have been on a steady decline with the rise of medical advancements and expanding education in treatment methods. Hospital death rates have experienced an overall decline from 2000 to 2010 with rates decreasing by 8%, despite the number of total hospitalizations increasing. This can be attributed to the progression of medical technology as well as scientific progress. Clinical investigations and the use of predictive modeling have led to these breakthroughs within the field and have ultimately increased life expectancy.

Using these models in our project to predict patient survival will help researchers identify which factors put patients at a higher risk of complication which will allow doctors and medical staff to recommend different treatment plans. This research will allow for future medical treatment to be more specific and tailored to each patient needs.

Acknowledgments

Agarwal, Mitisha. “Patient Survival Prediction.” Kaggle, 26 Dec. 2021, https://www.kaggle.com/datasets/mitishaagarwal/patient.

“Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 24 May 2017, www.cdc.gov/nchs/products/databriefs/db118.htm.