Introduction

The Problem Statement

I am working on this problem to help predict which patients will have a cardiac complication in the ICU. This will allow for proper utilization of hospital resources to identify which patients are at high risk. The criteria for success are to develop a model that has high recall and precision at identifying which patients will have cardiac complication. The scope of the solution will be to focus on identifying patients on admission who are at high risk for arrythmia, pulmonary edema, or death. Constraints on the scope will involve being able to get enough data on those patients that have a complication vs. those that do not. Key stakeholders will involve the hospitalist physicians in the Mid Atlantic Regional Health Center. The algorithm will be integrated and used in the EMR system. The key data source will be the UCI Machine Learning Repository. We combine the features in the table that contain the cardiac complication along with the feature denoting lethal outcome. The modeling response will be 1 for cardiac complication patients and 0 for patients without a cardiac complication. The models used in this project will be Deep Learning, Logistic Regression, Random Forest, and XGboost. The deliverables of the project will involve the Jupyter Notebooks for Data Wrangling, EDA, and Preprocessing Modeling. Also, a presentation slide deck, report, and metric report.

Background

The ability to predict a cardiac complication accurately is crucial for adjusting goals of care to the patients; for making sound medical decisions for management, treatment, and prevention.

Myocardial infarction is leading cause of death in most developed countries. The number of cases of heart attacks is one of leading cause of morbidity and mortality.

Predicting if someone with a Myocardial infarction will have a complication leading to an adverse outcome will help in the prevention of a lethal outcome. Given the prevalence of heart attacks and lethal outcomes, such a predictive algorithm would have the potential to save lives.

Identifying these high-risk patients and allocating the proper resources will decrease the number of cardiac arrests and rapid responses which are a considerate amount of stress for hospital staff.

Goals

This project aims to provide physicians with an identification of which patients are high risk for a cardiac complication. This will allow for the physician to allocate the proper level of care for high-risk patients and help prevent a lethal cardiac complication.

Datasets

The Dataset was downloaded from UCI Machine Learning Repository. The data was collected in the Krasnoyarsk Interdistrict Clinical Hospital No20 named after I. S. Berzon (Russia)

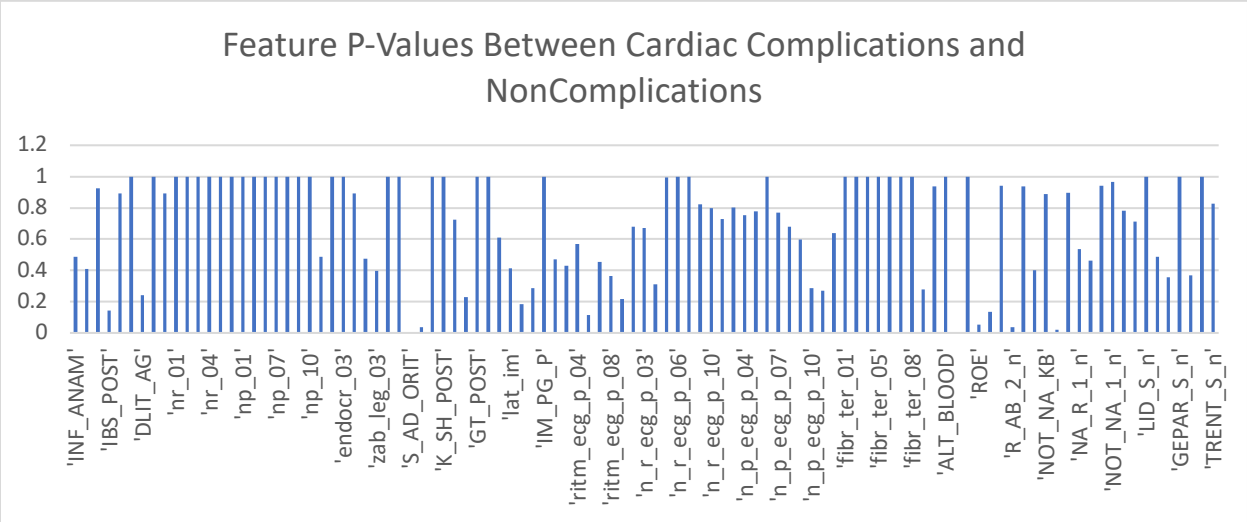
in 1992-1995. There was one csv file that contained all of the data. This dataset has 1700 patients with MI. The dataset has a total of 124 features. The last 12 features hold the complication and lethal outcome information. 7.6% of the data was NaN. For a complete set of the feature names and description, please see the Data Dictionary at the end of this document.

Data Wrangling and Feature Engineering

The dataset consisted of one CSV file. I proceeded to consolidate all the myocardial complications columns into one column. I then started to address all the NaN values. For the Age column, I calculated the mean age based on age and gender and used that value for the missing age column values. The 'SEX' column did not have any missing values. For the column 'IBS_NASL', I was only concerned with the patients that did indeed have hereditary CAD, and then consolidated values to either 0 and 1. 0 for not present, and 1 for present. I dropped the S_AD_KBRIG, and Ds_AD_KBRIG, due to the fact that over 90% of the values were missing. I then 'S_AD_ORIT', 'D_AD_ORIT' columns to a bin value between 0-4 based on the on American Heart Association guidelines for labeling blood pressure. I also binned the values of K_BLOOD, and NA_BLOOD based whether the values were normal, high, or low. I dropped the columns GIPER_NA, and GIPO_K because this information was captured in the K_BlOOD and NA_BLOOD. There were no duplicate values in the data. For all of the remaining columns with missing values I calculated the mean value based on age and gender and used that value for the missing column values.

Exploratory Data Analysis

For this part of the project, I went through feature by feature and did t-test if it was a numerical comparison or a chi squared if it was a categorical comparison. I was looking for statistical significance between the two populations. I used a p-value of <.05 to determine if there was statistical significance.



From this analysis only the following columns had a p-value of less than 0.05.

S_AD_ORIT
D_AD_ORIT
AST_BLOOD
R_AB_2_n
LID_KB

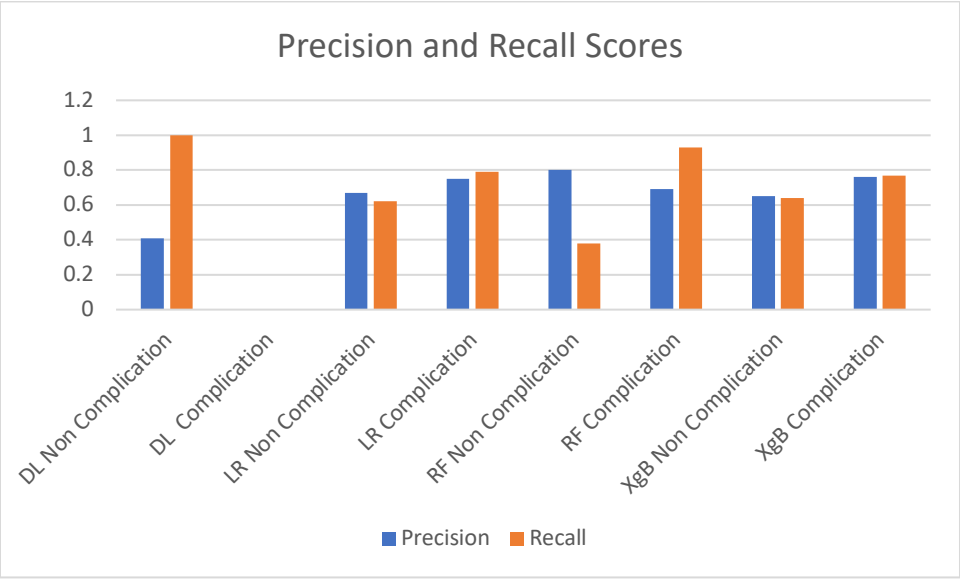
Model Description

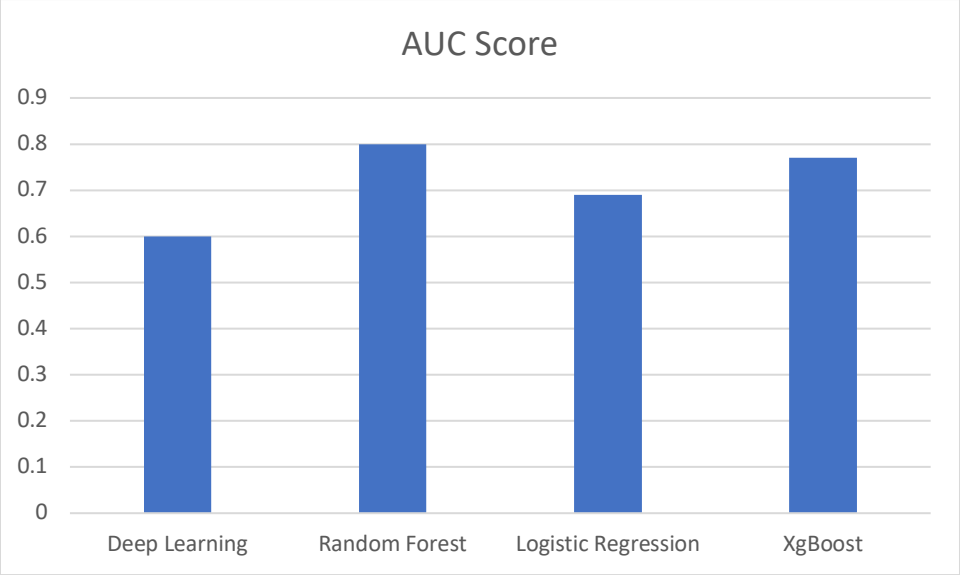
I trained four models. I used Deep Learning, Logistic Regression, Random Forrest, and XGboost. For the Deep Learning model I constructed a 2 layer sequential network. I used Relu activation for the first layer and a sigmoid activation for the output layer. When I compiled the model, I used 'sgd' optimizer and for loss I used categorical cross entropy. The model compiled with 20 epochs and the precision and recall scores for predicting complications was 0.

For Random Forest I did CV GridSearch to find optimized parameters. I found that the following variables were optimized as max depth=80, max features=3, min samples_leaf=3, min samples split= 8, and n estimators= 1000.

For Logistic Regression I did hyperparameter tuning on max iterations. I found the value of 100 to be optimal(train score = 0.617).

For XGboost I did hyperparameter tuning on Learning Rate and Estimator and found that the best learning rate was 0.5 and the best estimator was 500.





Model Findings

Using the Random Forest the model was able to predict with high degree of sensitivity and specificity which patients would likely have a cardiac complication. The Random Forest model would add value in the clinical setting. The precision and recall of this model are similar to the precision and recall to the strep urinary antigen test which is ubiquitous in medical practice.

The leading features in the RF model are the following:

- Age
- Presence of chronic HF
- History of Exertional angina
- Duration of arterial hypertension
- Gender
- Functional class (FC) of angina pectoris in the last year
- Presence of an essential hypertension
- History of Obstructive chronic bronchitis
- Coronary heart disease (CHD) in recent weeks, days before admission to hospital

For Logistic Regression are the top 12 Features Rank by Importance:

- LBBB on admission
- Type 1 Second-degree AV block on admission
- First-degree AV block
- Third-degree AV block
- Fibrinolytic therapy by Streptokinase
- Paroxysms of supraventricular tachycardia

- Ventricular fibrillation on ECG
- Use of opioid drugs in the ICU in the third day of the hospital period
- Ventricular fibrillation in PMH
- Cardiogenic shock at the time of admission to intensive care unit
- Relapse of the pain in the third day of the hospital period
- Presence of an inferior myocardial infarction

The features do align with a clinical assessment when determining the disease burden on a patient's heart.

Next Steps

For further research I would like to get a dataset from inpatient hospitalization that contains a more features and a larger patient population. I would also like to test on patients who do not have a myocardial infarction. I think looking at different cardiac outcomes such as which patients that present with chest pain will likely have coronary artery disease would be helpful.

Data Dictionary

Feature Name	Feature Description
ID	Record ID
AGE	Age
SEX	Gender
INF_ANAM	Quantity of myocardial infarctions in the anamnesis
STENOK_AN	Exertional angina pectoris in the anamnesis
FK_STENOK	Functional class (FC) of angina pectoris in the last year
IBS_POST	Coronary heart disease (CHD) in recent weeks, days before admission to hospital
IBS_NASL	Heredity on CHD
GB	Presence of essential hypertension
SIM_GIPERT	Symptomatic hypertension
DLIT_AG	Duration of arterial hypertension
ZSN_A	Presence of chronic Heart failure HF) in the anamnesis
nr11	Observing of arrhythmia in the anamnesis
nr01	Premature atrial contractions in the anamnesis
nr02	Premature ventricular contractions in the anamnesis
nr03	Paroxysms of atrial fibrillation in the anamnesis
nr04	A persistent form of atrial fibrillation in the anamnesis
nr07	Ventricular fibrillation in the anamnesis
nr08	Ventricular paroxysmal tachycardia in the anamnesis
np01	First-degree AV block in the anamnesis
np04	Third-degree AV block in the anamnesis
np05	LBBB in the anamnesis
np07	Incomplete LBBB in the anamnesis
np08	Complete LBBB in the anamnesis
np09	Incomplete RBBB in the anamnesis
np10	Complete RBBB in the anamnesis
endocr_01	Diabetes mellitus in the anamnesis
endocr_02	Obesity in the anamnesis
endocr_03	Thyrotoxicosis in the anamnesis
zab_leg_01	Chronic bronchitis in the anamnesis
zab_leg_02	Obstructive chronic bronchitis in the anamnesis
zab_leg_03	Bronchial asthma in the anamnesis
zab_leg_04	Chronic pneumonia in the anamnesis
zab_leg_06	Pulmonary tuberculosis in the anamnesis
S_AD_KBRIG	Systolic blood pressure according to Emergency Cardiology Team
D_AD_KBRIG	Diastolic blood pressure according to Emergency Cardiology Team
S_AD_ORIT	Systolic blood pressure according to intensive care unit
D_AD_ORIT	Diastolic blood pressure according to intensive care unit
O_L_POST	Pulmonary edema at the time of admission to intensive care unit
K_SH_POST	Cardiogenic shock at the time of admission to intensive care unit
MP_TP_POST	Paroxysms of atrial fibrillation at the time of admission to intensive care unit, or at a pre-hospital stage
SVT_POST	Paroxysms of supraventricular tachycardia at the time of admission to intensive care unit, or at a pre-hospital stage
GT_POST	Paroxysms of ventricular tachycardia at the time of admission to intensive care unit, or at a pre-hospital stage
FIB_G_POST	Ventricular fibrillation at the time of admission to intensive care unit, or at a pre-hospital stage
ant_im	Presence of an anterior myocardial infarction (left ventricular)
lat_im	Presence of a lateral myocardial infarction (left ventricular)
inf_im	Presence of an inferior myocardial infarction (left ventricular)
post_im	Presence of a posterior myocardial infarction (left ventricular)
IM_PG_P	Presence of a right ventricular myocardial infarction

ritm_ecg_p_01	ECG rhythm at the time of admission to hospital – sinus with a heart rate 60-90
ritm_ecg_p_02	ECG rhythm at the time of admission to hospital – atrial fibrillation
ritm_ecg_p_04	ECG rhythm at the time of admission to hospital – atrial
ritm_ecg_p_06	ECG rhythm at the time of admission to hospital – idioventricular
ritm_ecg_p_07	ECG rhythm at the time of admission to hospital – sinus with a heart rate above 90 (tachycardia)
ritm_ecg_p_08	ECG rhythm at the time of admission to hospital – sinus with a heart rate below 60 (bradycardia)
n_r_ecg_p_01	Premature atrial contractions on ECG at the time of admission to hospital
n_r_ecg_p_02	Frequent premature atrial contractions on ECG at the time of admission to hospital
n_r_ecg_p_03	Premature ventricular contractions on ECG at the time of admission to hospital
n_r_ecg_p_04	Frequent premature ventricular contractions on ECG at the time of admission to hospital
n_r_ecg_p_05	Paroxysms of atrial fibrillation on ECG at the time of admission to hospital
n_r_ecg_p_06	Persistent form of atrial fibrillation on ECG at the time of admission to hospital
n_r_ecg_p_08	Paroxysms of supraventricular tachycardia on ECG at the time of admission to hospital
n_r_ecg_p_09	Paroxysms of ventricular tachycardia on ECG at the time of admission to hospital
n_r_ecg_p_10	Ventricular fibrillation on ECG at the time of admission to hospital
n_p_ecg_p_01	Sinoatrial block on ECG at the time of admission to hospital
n_p_ecg_p_03	First-degree AV block on ECG at the time of admission to hospital
n_p_ecg_p_04	Type 1 Second-degree AV block (Mobitz I/Wenckebach) on ECG at the time of admission to hospital
n_p_ecg_p_05	Type 2 Second-degree AV block (Mobitz II/Hay) on ECG at the time of admission to hospital
n_p_ecg_p_06	Third-degree AV block on ECG at the time of admission to hospital
n_p_ecg_p_07	LBBB (anterior branch) on ECG at the time of admission to hospital
n_p_ecg_p_08	LBBB (posterior branch) on ECG at the time of admission to hospital
n_p_ecg_p_09	Incomplete LBBB on ECG at the time of admission to hospital
n_p_ecg_p_10	Complete LBBB on ECG at the time of admission to hospital
n_p_ecg_p_11	Incomplete RBBB on ECG at the time of admission to hospital
n_p_ecg_p_12	Complete RBBB on ECG at the time of admission to hospital
fibr_ter_01	Fibrinolytic therapy by Celiasum 750k IU
fibr_ter_02	Fibrinolytic therapy by Celiasum 1m IU
fibr_ter_03	Fibrinolytic therapy by Celiasum 3m IU
fibr_ter_05	Fibrinolytic therapy by Streptase
fibr_ter_06	Fibrinolytic therapy by Celiasum 500k IU
fibr_ter_07	Fibrinolytic therapy by Celiasum 250k IU
fibr_ter_08	Fibrinolytic therapy by Streptodecase 1.5m IU
GIPO_K	Hypokalemia (< 4 mmol/L)
K_BLOOD	Serum potassium content
GIPER_Na	Increase of sodium in serum (more than 150 mmol/L)
Na_BLOOD	Serum sodium content
ALT_BLOOD	Serum AlAT content
AST_BLOOD	Serum AsAT content
KFK_BLOOD	Serum CPK content
L_BLOOD	White blood cell count (billions per liter)
ROE	ESR (Erythrocyte sedimentation rate)
TIME_B_S	Time elapsed from the beginning of the attack of CHD to the hospital
R_AB_1_n	Relapse of the pain in the first hours of the hospital period
R_AB_2_n	Relapse of the pain in the second day of the hospital period
R_AB_3_n	Relapse of the pain in the third day of the hospital period
NA_KB	Use of opioid drugs by the Emergency Cardiology Team
NOT_NA_KB	Use of NSAIDs by the Emergency Cardiology Team
LID_KB	Use of lidocaine by the Emergency Cardiology Team
NITR_S	Use of liquid nitrates in the ICU

NA_R_1_n	Use of opioid drugs in the ICU in the first hours of the hospital period
NA_R_2_n	Use of opioid drugs in the ICU in the second day of the hospital period
NA_R_3_n	Use of opioid drugs in the ICU in the third day of the hospital period
NOT_NA_1_n	Use of NSAIDs in the ICU in the first hours of the hospital period
NOT_NA_2_n	Use of NSAIDs in the ICU in the second day of the hospital period
NOT_NA_3_n	Use of NSAIDs in the ICU in the third day of the hospital period
LID_S_n	Use of lidocaine in the ICU
B_BLOK_S_n	Use of beta-blockers in the ICU
ANT_CA_S_n	Use of calcium channel blockers in the ICU
GEPAR_S_n	Use of a anticoagulants (heparin) in the ICU
ASP_S_n	Use of acetylsalicylic acid in the ICU
TIKL_S_n	Use of Ticlid in the ICU
TRENT_S_n	Use of Trental in the ICU
FIBR_PREDS	Atrial fibrillation
PREDS_TAH	Supraventricular tachycardia
JELUD_TAH	Ventricular tachycardia
FIBR_JELUD	Ventricular fibrillation
A_V_BLOK	Third-degree AV block
OTEK_LANC	Pulmonary edema
RAZRIV	Myocardial rupture
DRESSLER	Dressler syndrome
ZSN	Chronic heart failure
REC_IM	Relapse of the myocardial infarction
P_IM_STEN	Post-infarction angina
LET_IS	Lethal outcome (cause)