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.

Feature Description:

1. Record ID (ID).

2. Age (AGE).

3. Gender (SEX):
0 – female
1 – male

4. Quantity of myocardial infarctions in the anamnesis (INF_ANAM):
0 – zero
1 – one
2 – two
3 – three and more

5. Exertional angina pectoris in the anamnesis (STENOK_AN): 0 – never
1 – during the last year 2 – one year ago
3 – two years ago
4 – three years ago
5 – 4-5 years ago
6 – more than 5 years ago

6. Functional class (FC) of angina pectoris in the last year (FK_STENOK)[2]:
0 – there is no angina pectoris 1 – I FC
2 – II FC
3 – III FC.
4 – IV FC

7. Coronary heart disease (CHD) in recent weeks, days before admission to hospital (IBS_POST):
0 – there was no CHD
1 – exertional angina pectoris 2 – unstable angina pectoris

8. Heredity on CHD (IBS_NASL): 0 – isn't burdened
1 – burdened

9. Presence of an essential hypertension (GB):
0 – there is no essential hypertension 1 – Stage 1
2 – Stage 2
3 – Stage

10. Symptomatic hypertension (SIM_GIPERT):
0 – no
1 – yes

11. Duration of arterial hypertension (DLIT_AG):
0 – there was no arterial hypertension 1 – one year
2 – two years
3 – three years
4 – four years
5 – five years
6 – 6-10 years
7 – more than 10 years

12. Presence of chronic Heart failure (HF) in the anamnesis (ZSN_A):
0 – there is no chronic heart failure
1 – I stage
2 – IIA stage (heart failure due to right ventricular systolic dysfunction)
3 – IIA stage (heart failure due to left ventricular systolic dysfunction)
4 – IIB stage (heart failure due to left and right ventricular systolic dysfunction) 5 – III stage (dystrophic changes in organs)

13. Observing of arrhythmia in the anamnesis (nr11): 0 – no
1 – yes

14. Premature atrial contractions in the anamnesis (nr01):
0 – no
1 – yes

15. Premature ventricular contractions in the anamnesis (nr02):
0 – no
1 – yes

16. Paroxysms of atrial fibrillation in the anamnesis (nr03):
0 – no
1 – yes

17. A persistent form of atrial fibrillation in the anamnesis (nr04):
0 – no
1 – yes

18. Ventricular fibrillation in the anamnesis (nr07):
0 – no
1 – yes

19. Ventricular paroxysmal tachycardia in the anamnesis (nr08):
0 – no
1 – yes

20. First-degree AV block in the anamnesis (np01):
0 – no
1 – yes

21. Third-degree AV block in the anamnesis (np04):
0 – no
1 – yes

22. LBBB (anterior branch) in the anamnesis (np05):
0 – no
1 – yes

23. Incomplete LBBB in the anamnesis (np07):
0 – no
1 – yes

24. Complete LBBB in the anamnesis (np08):
0 – no
1 – yes

25. Incomplete RBBB in the anamnesis (np09):
0 – no
1 – yes

26. Complete RBBB in the anamnesis (np10):
0 – no
1 – yes

27. Diabetes mellitus in the anamnesis (endocr_01):
0 – no
1 – yes

28. Obesity in the anamnesis (endocr_02):
0 – no
1 – yes

29. Thyrotoxicosis in the anamnesis (endocr_03):
0 – no
1 – yes

30. Chronic bronchitis in the anamnesis (zab_leg_01):
0 – no
1 – yes

31.Obstructive chronic bronchitis in the anamnesis (zab_leg_02):
0 – no
1 – yes

32. Bronchial asthma in the anamnesis (zab_leg_03):
0 – no
1 – yes

33. Chronic pneumonia in the anamnesis (zab_leg_04):
0 – no
1 – yes

34. Pulmonary tuberculosis in the anamnesis (zab_leg_06):
0 – no
1 – yes

35. Systolic blood pressure according to Emergency Cardiology Team (S_AD_KBRIG) (mmHg).

36. Diastolic blood pressure according to Emergency Cardiology Team (D_AD_KBRIG) (mmHg).

37. Systolic blood pressure according to intensive care unit (S_AD_ORIT) (mmHg).

38. Diastolic blood pressure according to intensive care unit (D_AD_ORIT) (mmHg).

39. Pulmonary edema at the time of admission to intensive care unit (O_L_POST):
0 – no
1 – yes

40. Cardiogenic shock at the time of admission to intensive care unit (K_SH_POST):
0 – no
1 – yes

41. Paroxysms of atrial fibrillation at the time of admission to intensive care unit, (or at a pre-hospital stage) (MP_TP_POST):
0 – no
1 – yes

42. Paroxysms of supraventricular tachycardia at the time of admission to intensive care unit, (or at a pre-hospital stage) (SVT_POST):
0 – no 1 – yes

43. Paroxysms of ventricular tachycardia at the time of admission to intensive care unit, (or at a pre-hospital stage) (GT_POST):
0 – no
1 – yes

44. Ventricular fibrillation at the time of admission to intensive care unit, (or at a pre-hospital stage) (FIB_G_POST):
0 – no
1 – yes

45. Presence of an anterior myocardial infarction (left ventricular) (ECG changes in leads V1 – V4) (ant_im):
0 – no
1 – QRS has no changes
2 – QRS is like QR-complex 3 – QRS is like Qr-complex 4 – QRS is like QS-complex

46. Presence of a lateral myocardial infarction (left ventricular) (ECG changes in leads V5 – V6 , I, AVL) (lat_im):
0 – no
1 – QRS has no changes
2 – QRS is like QR-complex 3 – QRS is like Qr-complex 4 – QRS is like QS-complex

47. Presence of an inferior myocardial infarction (left ventricular) (ECG changes in leads III, AVF, II). (inf_im):
0 – no
1 – QRS has no changes
2 – QRS is like QR-complex 3 – QRS is like Qr-complex 4 – QRS is like QS-complex

48. Presence of a posterior myocardial infarction (left ventricular) (ECG changes in V7 – V9, reciprocity changes in leads V1 – V3) (post_im):
0 – no
1 – QRS has no changes
2 – QRS is like QR-complex 3 – QRS is like Qr-complex 4 – QRS is like QS-complex

49. Presence of a right ventricular myocardial infarction (IM_PG_P): 0 – no
1 – yes
50. ECG rhythm at the time of admission to hospital – sinus (with a heart rate 60-90) (ritm_ecg_p_01):
0 – no
1 – yes

51. ECG rhythm at the time of admission to hospital – atrial fibrillation (ritm_ecg_p_02):
0 – no
1 – yes

52. ECG rhythm at the time of admission to hospital – atrial (ritm_ecg_p_04):
0 – no
1 – yes

53. ECG rhythm at the time of admission to hospital – idioventricular (ritm_ecg_p_06):
0 – no 1 – yes

54. ECG rhythm at the time of admission to hospital – sinus with a heart rate above 90 (tachycardia) (ritm_ecg_p_07):
0 – no
1 – yes

55. ECG rhythm at the time of admission to hospital – sinus with a heart rate below 60 (bradycardia) (ritm_ecg_p_08):
0 – no
1 – yes

56. Premature atrial contractions on ECG at the time of admission to hospital (n_r_ecg_p_01):
0 – no
1 – yes

57. Frequent premature atrial contractions on ECG at the time of admission to hospital (n_r_ecg_p_02):
0 – no
1 – yes

58.Premature ventricular contractions on ECG at the time of admission to hospital (n_r_ecg_p_03):
0 – no
1 – yes

59. Frequent premature ventricular contractions on ECG at the time of admission to hospital (n_r_ecg_p_04):
0 – no
1 – yes

60. Paroxysms of atrial fibrillation on ECG at the time of admission to hospital (n_r_ecg_p_05):
0 – no
1 – yes

61. Persistent form of atrial fibrillation on ECG at the time of admission to hospital (n_r_ecg_p_06):
0 – no
1 – yes

62. Paroxysms of supraventricular tachycardia on ECG at the time of admission to hospital (n_r_ecg_p_08):
0 – no
1 – yes

63. Paroxysms of ventricular tachycardia on ECG at the time of admission to hospital (n_r_ecg_p_09):
0 – no
1 – yes

64. Ventricular fibrillation on ECG at the time of admission to hospital (n_r_ecg_p_10):
0 – no
1 – yes

65. Sinoatrial block on ECG at the time of admission to hospital (n_p_ecg_p_01):
0 – no
1 – yes

66. First-degree AV block on ECG at the time of admission to hospital (n_p_ecg_p_03):
0 – no
1 – yes

67. Type 1 Second-degree AV block (Mobitz I/Wenckebach) on ECG at the time of admission to hospital (n_p_ecg_p_04):
0 – no 1 – yes

68. Type 2 Second-degree AV block (Mobitz II/Hay) on ECG at the time of admission to hospital (n_p_ecg_p_05):
0 – no
1 – yes

69. Third-degree AV block on ECG at the time of admission to hospital (n_p_ecg_p_06):
0 – no
1 – yes

70. LBBB (anterior branch) on ECG at the time of admission to hospital (n_p_ecg_p_07):
0 – no
1 – yes

71. LBBB (posterior branch) on ECG at the time of admission to hospital (n_p_ecg_p_08):
0 – no
1 – yes

72. Incomplete LBBB on ECG at the time of admission to hospital (n_p_ecg_p_09):
0 – no
1 – yes

73. Complete LBBB on ECG at the time of admission to hospital (n_p_ecg_p_10):
0 – no
1 – yes

74. Incomplete RBBB on ECG at the time of admission to hospital (n_p_ecg_p_11):
0 – no
1 – yes

75. Complete RBBB on ECG at the time of admission to hospital (n_p_ecg_p_12):
0 – no
1 – yes

76. Fibrinolytic therapy by Celiasum 750k IU (fibr_ter_01):
0 – no
1 – yes

77. Fibrinolytic therapy by Celiasum 1m IU (fibr_ter_02):
0 – no
1 – yes

78. Fibrinolytic therapy by Celiasum 3m IU (fibr_ter_03):
0 – no
1 – yes

79. Fibrinolytic therapy by Streptase (fibr_ter_05):
0 – no
1 – yes

80. Fibrinolytic therapy by Celiasum 500k IU (fibr_ter_06):
0 – no
1 – yes

81. Fibrinolytic therapy by Celiasum 250k IU (fibr_ter_07):
0 – no
1 – yes

82. Fibrinolytic therapy by Streptodecase 1.5m IU (fibr_ter_08):
0 – no
1 – yes

83. Hypokalemia (< 4 mmol/L) (GIPO_K):
0 – no
1 – yes

84. Serum potassium content (K_BLOOD) (mmol/L).

85 Increase of sodium in serum (more than 150 mmol/L) (GIPER_Na):
0 – no
1 – yes

86. Serum sodium content (Na_BLOOD) (mmol/L).

87. Serum AlAT content (ALT_BLOOD) (IU/L).

88. Serum AsAT content (AST_BLOOD) (IU/L).

89. Serum CPK content (KFK_BLOOD) (IU/L).

90. White blood cell count (billions per liter) (L_BLOOD).

91. ESR (Erythrocyte sedimentation rate) (ROE) (мм).

92. Time elapsed from the beginning of the attack of CHD to the hospital (TIME_B_S):
1 – less than 2 hours 2 – 2-4 hours
3 – 4-6 hours
4 – 6-8 hours
5 – 8-12 hours
6 – 12-24 hours
7 – more than 1 days 8 – more than 2 days 9 – more than 3 days

93. Relapse of the pain in the first hours of the hospital period (R_AB_1_n): 0 – there is no relapse
1 – only one
2 – 2 times
3 – 3 or more times

94. Relapse of the pain in the second day of the hospital period (R_AB_2_n): 0 – there is no relapse
1 – only one
2 – 2 times
3 – 3 or more times

95. Relapse of the pain in the third day of the hospital period (R_AB_3_n): 0 – there is no relapse
1 – only one
2 – 2 times
3 – 3 or more times

96. Use of opioid drugs by the Emergency Cardiology Team (NA_KB): 0 – no
1 – yes

97. Use of NSAIDs by the Emergency Cardiology Team (NOT_NA_KB):
0 – no
1 – yes

98.Use of lidocaine by the Emergency Cardiology Team (LID_KB):
0 – no
1 – yes

99. Use of liquid nitrates in the ICU (NITR_S):
0 – no
1 – yes

100. Use of opioid drugs in the ICU in the first hours of the hospital period (NA_R_1_n):
0 – no
1 – once
2 – twice
3 – three times 4 – four times

101. Use of opioid drugs in the ICU in the second day of the hospital period (NA_R_2_n): 0 – no

1 – once
2 – twice
3 – three times 4 – four times

102. Use of opioid drugs in the ICU in the third day of the hospital period (NA_R_3_n): 0 – no
1 – once
2 – twice
3 – three times 4 – four times

103. Use of NSAIDs in the ICU in the first hours of the hospital period (NOT_NA_1_n): 0 – no
1 – once
2 – twice
3 – three times
4 – four or more times

104. Use of NSAIDs in the ICU in the second day of the hospital period (NOT_NA_2_n): 0 – no
1 – once
2 – twice
3 – three times
4 – four or more times

105. Use of NSAIDs in the ICU in the third day of the hospital period (NOT_NA_3_n): 0 – no
1 – once
2 – twice
3 – three times
4 – four or more times

106. Use of lidocaine in the ICU (LID_S_n): 0 – no
1 – yes

107. Use of beta-blockers in the ICU (B_BLOK_S_n):
0 – no
1 – yes

108. Use of calcium channel blockers in the ICU (ANT_CA_S_n):
0 – no
1 – yes

109. Use of a anticoagulants (heparin) in the ICU (GEPAR_S_n):
0 – no
1 – yes

110. Use of acetylsalicylic acid in the ICU (ASP_S_n):
0 – no
1 – yes

111. Use of Ticlid in the ICU (TIKL_S_n):
0 – no
1 – yes

112. Use of Trental in the ICU (TRENT_S_n):
0 – no 1 – yes

These are the Target Features that were combined into one Feature:

113. Atrial fibrillation (FIBR_PREDS): 0 – no
1 – yes

114. Supraventricular tachycardia (PREDS_TAH):
0 – no
1 – yes

115. Ventricular tachycardia (JELUD_TAH):
0 – no
1 – yes

116. Ventricular fibrillation (FIBR_JELUD):
0 – no
1 – yes

117. Third-degree AV block (A_V_BLOK):
0 – no
1 – yes

118. Pulmonary edema (OTEK_LANC):
0 – no
1 – yes

119. Myocardial rupture (RAZRIV):
0 – no
1 – yes

120. Dressler syndrome (DRESSLER):
0 – no
1 – yes

121. Chronic heart failure (ZSN):
0 – no
1 – yes

122. Relapse of the myocardial infarction (REC_IM):
0 – no
1 – yes

123. Post-infarction angina (P_IM_STEN):
0 – no
1 – yes

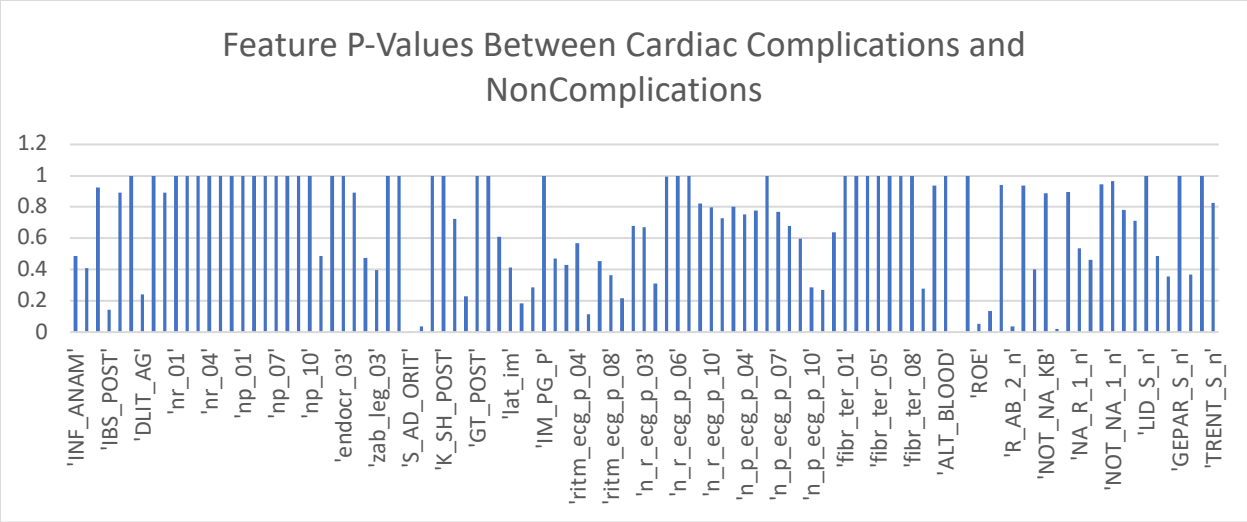
124. Lethal outcome (cause) (LET_IS):
0 – unknown
1 – cardiogenic shock
2 – pulmonary edema
3 – myocardial rupture
4 – progress of congestive heart failure 5 – thromboembolism
6 – asystole
7 – ventricular fibrillation

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

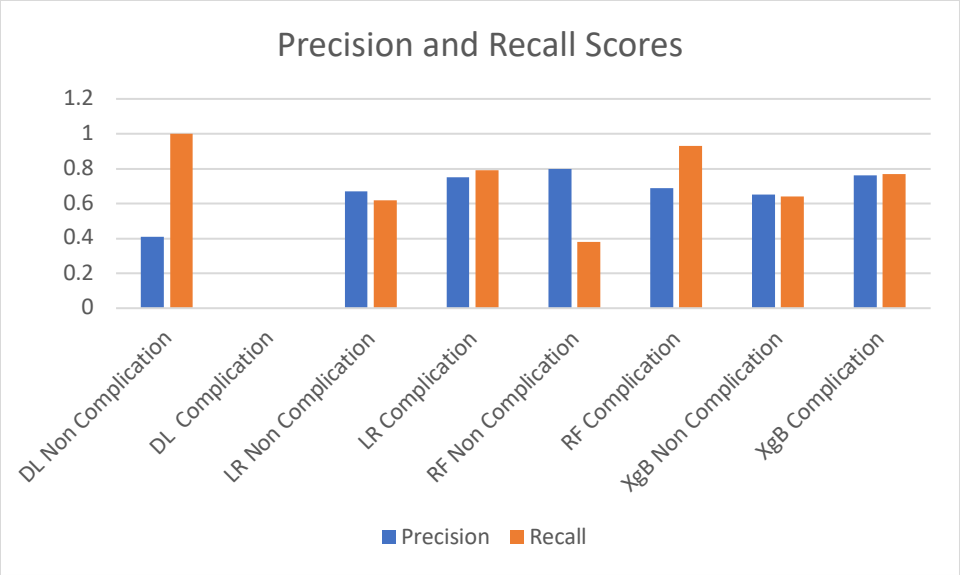
Model Description

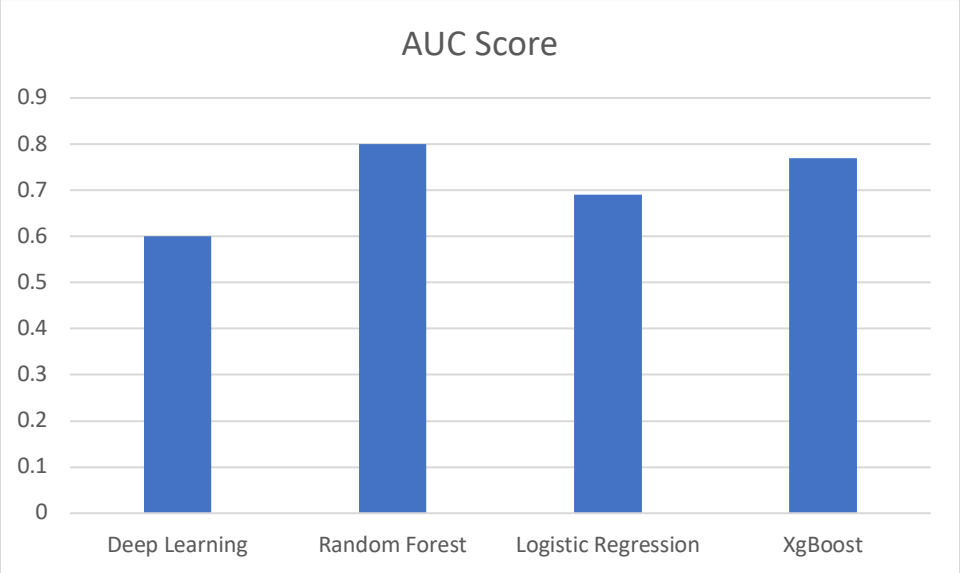
I trained four models. I used Deep Learning, Logistic Regression, Random Forrest, and XGboost. For the Deep Learning model I constructed a 2 layer sequential network. I used Relu activation for the first layer and a sigmoid activation for the output layer. When I compiled the model, I used 'sgd' optimizer 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.