CSCI 381/780 – Applied Data Science
Final Report
Tao Hu

Project Abstract

Myocardial infarction (MI) is one of the dangerous diseases. The course of the disease in patients with MI is different. MI can occur without complications or with complications that do not worsen the long-term prognosis. At the same time, about half of patients in the acute and subacute periods have complications leading to a worsening of the course of the disease and even death.

The aims of this project were to create a ML model that can classify whether a patient has one of myocardial infarction complications Myocardial rupture (RAZRIV) Based on record in the patient's medical history, I approach this goal by use model Decision Tree, Random Forest, Support vector machines (SVM), Multi-layer Perceptron. The Random Forest model performed the best, with sensitivity score of 0.9832 and precision 0.54 However, all models faced a limitation in their ability to predict positive case of Myocardial rupture due to extremely imbalanced dataset, getting more positive cases record will be the key to improving model performance. Due to data limitations, I will consider different thresholds and tradeoffs of precision between negative and positive case for MLP model.

Accomplishments

The main research goal of this project was to construct a model to classify whether a patient has Myocardial rupture base on patient records and to determine the heavy associated cause to help doctors make better judgments

Specific Objectives

I completed all of Aim 1, the dataset contains large numbers of missing value, filled with mean by imputation method, split data into training data, test data and validation data, determined RAZRIV as target feature, feature engineering and feature selection.

I completed all of Aim 2, which is evaluate different model and determine the best model with highest AUC and precision score for positive case, also my model can predict possibility of patient have Myocardial rupture by record of the time of admission to hospital (Day 1)

[1]This template was adapted from the standard NSF research progress report (RPPR).

Major Activities

In data processing, I dropped irrelevant column **ID** from dataset, and determined column 118 Myocardial rupture (RAZRIV) as target(predict) feature, use seaborn to plot a graph to check number of positive and negative case number in dataset and observe target feature distribution on input features.

In data splitting, the dataset was split into a 60% training set, 20% validation set, and 20% test set.

In Decision Tree model, theses' overfitting, I applied k-fold cross validation to training set, The case is very much skewed, Myocardial rupture have 1646 negative case and 54 positive cases, if do a simple k-fold, we will not have an equal distribution of targets in every fold. Thus, I choose stratified k-fold cross validation in this case, plot a graph at end, the model performance is bad as expected.

In Random Forest and SVM, process Hyperparameter Tuning, I set scoring to precision in Randomized Search CV as precision metrics is one of most important metrics for this project and save parameter for feature selection

In feature selection, evaluate Random Forest and SVM trained in Hyperparameter Tuning, calculate performance 5 times with most relevant features, corresponding to top 2 features, top 5 features, top 10 features, top 20 features, top 80 features.

In MLP, Normalize the input features using the sklearn StandardScaler. This will set the mean to 0 and standard deviation to 1. Set the correct initial bias, and the model will give much more reasonable initial guesses. Define callback functions monitor precision metrics, it helps reduce training time, evaluate model performance, then apply model with threshold, model with weighted class, model with oversampling, model with threshold and weighted class.

Significant Findings

Label distribution and missingness analysis:

[1]This template was adapted from the standard NSF research progress report (RPPR).

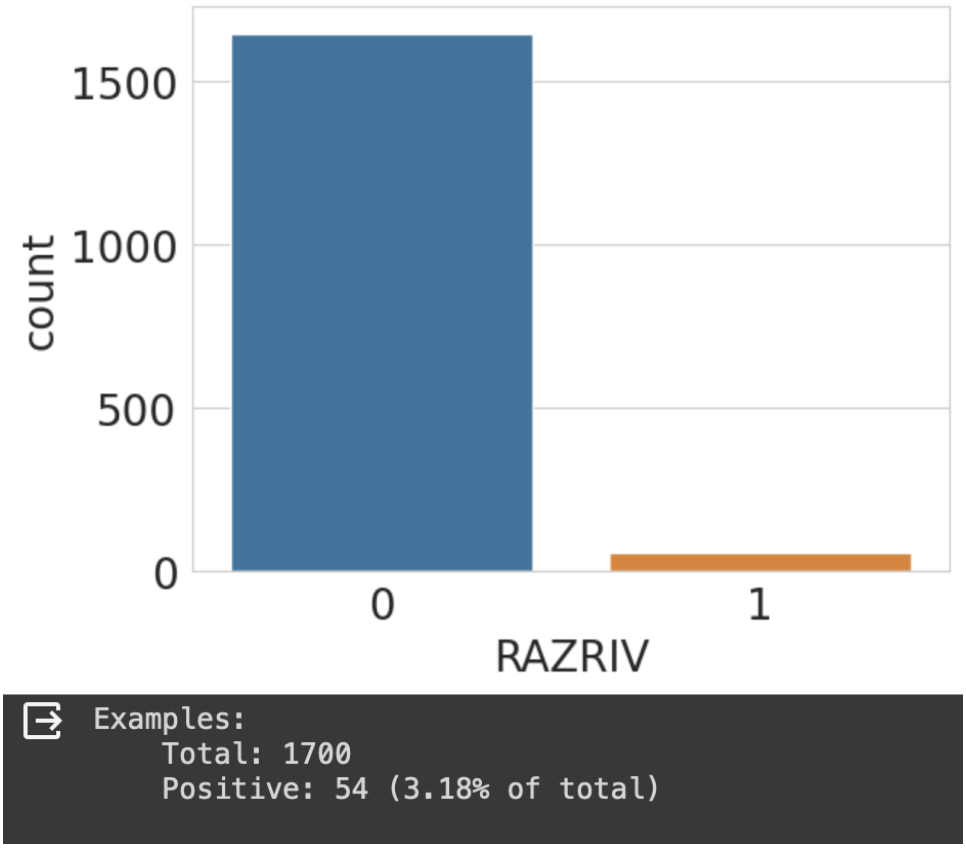

Figure 1

Investigating the distribution of status' found the sampled data shows an uneven distribution with only 3.18% of observations having status 1 (positive case). This is imbalanced data

Overfitting:

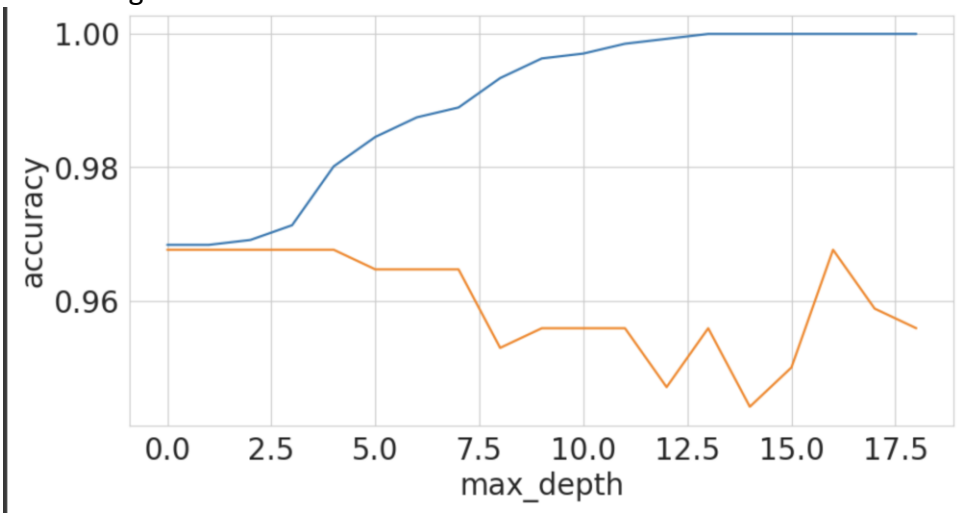

Figure 2

[1]This template was adapted from the standard NSF research progress report (RPPR).

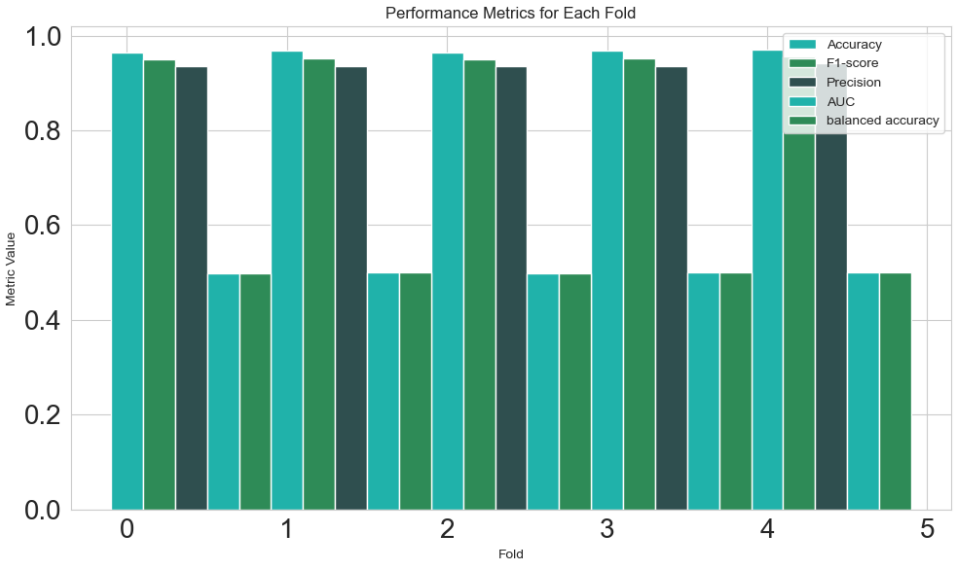

Figure 3

```
Fold: 0, Accuracy: 0.9647058823529412
Fold: 0, f1: 0.9502641775272983
Fold: 0, precision: 0.9362484816935623
Fold: 0, AUC: 0.49848024316109424
Fold: 0, balanced accuracy: 0.49848024316109424
Confusion Matrix:
 [[328   1]
 [ 11   0]]
Fold: 1, Accuracy: 0.9676470588235294
Fold: 1, f1: 0.9517365690670886
Fold: 1, precision: 0.9363408304498271
Fold: 1, AUC: 0.5
Fold: 1, balanced accuracy: 0.5
Confusion Matrix:
 [[329   0]
 [ 11   0]]
Fold: 2, Accuracy: 0.9647058823529412
Fold: 2, f1: 0.9502641775272983
Fold: 2, precision: 0.9362484816935623
Fold: 2, AUC: 0.49848024316109424
Fold: 2, balanced accuracy: 0.49848024316109424
Confusion Matrix:
 [[328   1]
 [ 11   0]]
Fold: 3, Accuracy: 0.9676470588235294
Fold: 3, f1: 0.9517365690670886
Fold: 3, precision: 0.9363408304498271
Fold: 3, AUC: 0.5
Fold: 3, balanced accuracy: 0.5
Confusion Matrix:
 [[329   0]
 [ 11   0]]
Fold: 4, Accuracy: 0.9705882352941176
Fold: 4, f1: 0.9561018437225636
Fold: 4, precision: 0.9420415224913495
Fold: 4, AUC: 0.5
Fold: 4, balanced accuracy: 0.5
Confusion Matrix:
 [[330   0]
 [ 10   0]]
```

Figure 4

[1]This template was adapted from the standard NSF research progress report (RPPR).

The decision Tree experience overfitting during training after depth 3, after stratified k-fold cross validation, we get consistence result, we get accuracy about 95% for each folder, notice balanced accuracy and position of matrix [1][1], balanced accuracy is around 50% and model cannot recognize positive case, there are 11 incorrect predictions and 0 correct predictions, in practical, improving the model's recognition of patient having Myocardial rupture has significant implications, accuracy is not a proper metrics for this project, we cannot use default metrics to calculate accuracy, f1, precision, AUC from scikit-learn, we need define these metric manually.

- **Precision** is the percentage of **predicted** positives that were correctly classified > $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$
- **Recall** is the percentage of **actual** positives that were correctly classified > $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

Figure 5

SVM:

```
Best Hyperparameters: {'C': 0.005764194027503303, 'degree': 8, 'gamma': 41.91464667297158, 'kernel': 'linear'}
Best Precision Score: 0.3347962382445141
```

Figure 6

Random Forest:

```
Best Hyperparameters: {'bootstrap': True, 'max_depth': 57, 'max_features': 'log2', 'min_samples_leaf': 0.014336132600544057, 'm
in_samples_split': 0.0490005896214099, 'n_estimators': 801}
Best precision Score: 0.4821428571428571
```

Figure 7

Feature Selection:

```
Top 2 Features:
SVM Metrics (Validation Set):
confusion_matrix:
[[312  15]
 [  0  13]]
Random Forest Metrics (Validation Set):
confusion_matrix:
[[316  11]
 [  0  13]]
```

[1]This template was adapted from the standard NSF research progress report (RPPR).

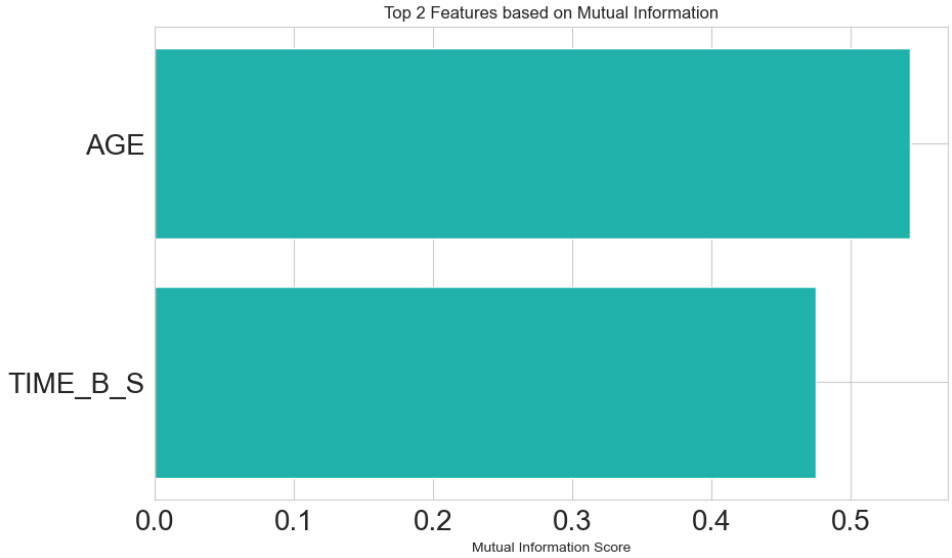

Figure 8

```
Top 5 Features:
SVM Metrics (Validation Set):
confusion_matrix:
[[312  15]
 [  0  13]]
Random Forest Metrics (Validation Set):
confusion_matrix:
[[316  11]
 [  0  13]]
```

Figure 9

```
Top 10 Features:
SVM Metrics (Validation Set):
confusion_matrix:
[[312  15]
```

[1]This template was adapted from the standard NSF research progress report (RPPR).

```
 [  0  13]]
Random Forest Metrics (Validation Set):

confusion_matrix:
[[315  12]
 [  0  13]]
```

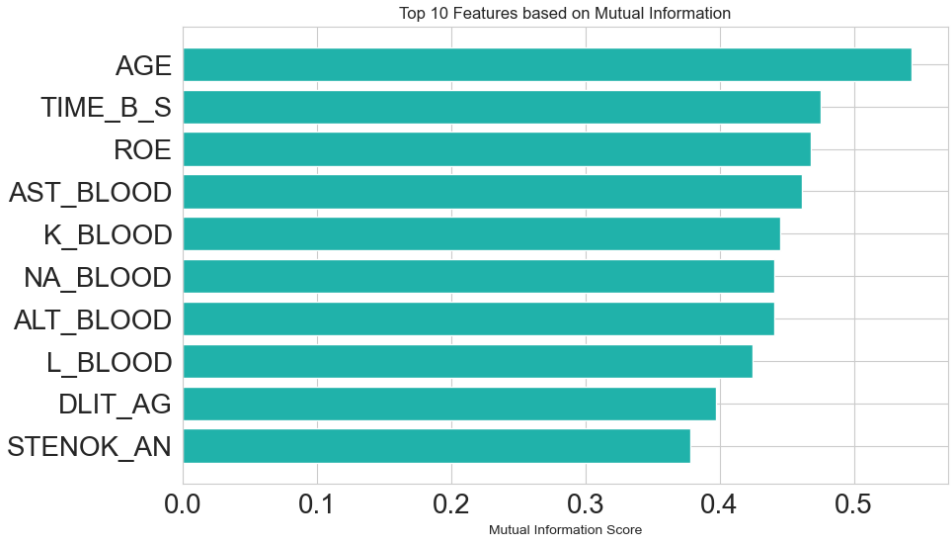

Figure 10

```
Top 20 Features:
SVM Metrics (Validation Set):
confusion_matrix:
[[312  15]
 [  0  13]]
Random Forest Metrics (Validation Set):
confusion_matrix:
[[318   9]
 [  1  12]]
```

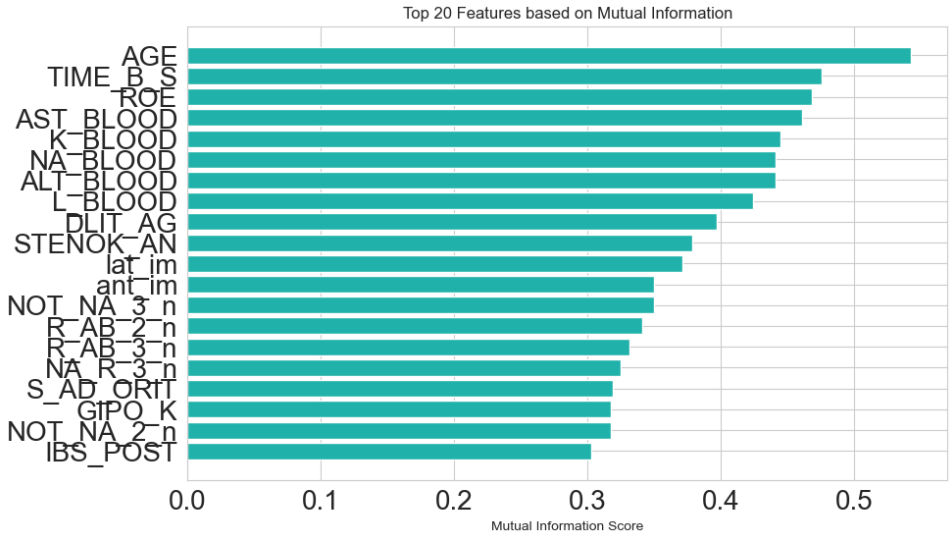

Figure 11

[1]This template was adapted from the standard NSF research progress report (RPPR).

In Feature Selection, we can find that at top 5 features are:
AGE: Age
TIME_B_S: Time elapsed from the beginning of the attack of CHD to the hospital
ROE: ESR (Erythrocyte sedimentation rate):
AST_BLOOD: Serum AsAT content
K_BLOOD: Serum potassium content

we get best precision and sensitivity, all true positive case recognized by the model , due to lack of positive case data, but we can confirm that these 5 non-complications cause Myocardial rupture

P.S. In experiment one of complications post-infarction angina (P_IM_STEN) is heavy associated cause for Myocardial rupture (RAZRIV), but it is predication label and one of complications

Case Distribution

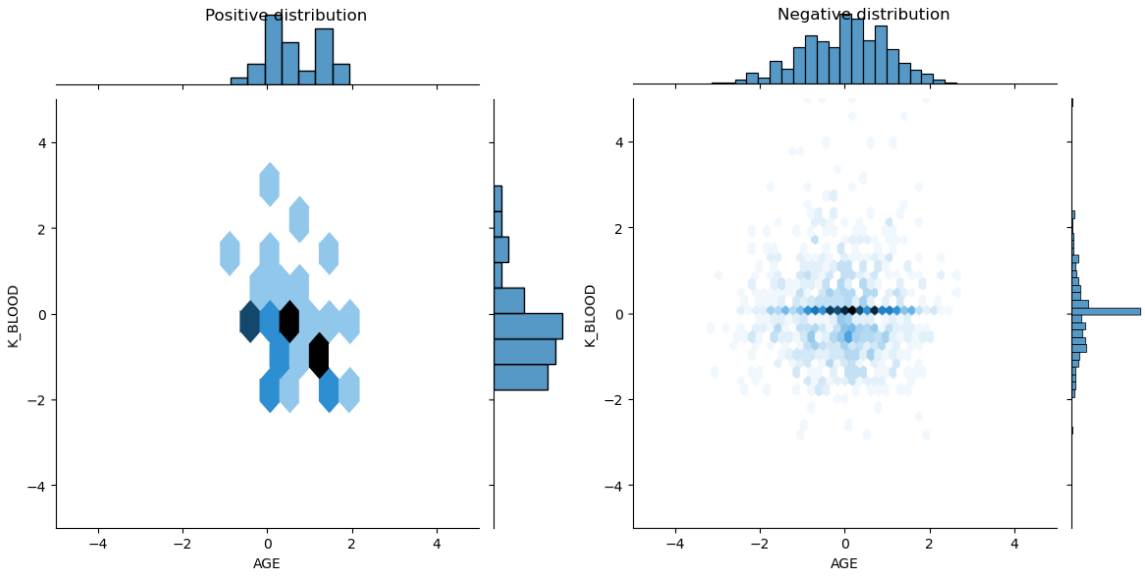

Figure 12

Once we confirm the most relevant feature to our target feature Myocardial rupture (RAZRIV), I plot a graph to observe the case distribution of these top features, deeper colors in the chart represent higher numbers of cases, and vice versa.

MLP:

[1]This template was adapted from the standard NSF research progress report (RPPR).

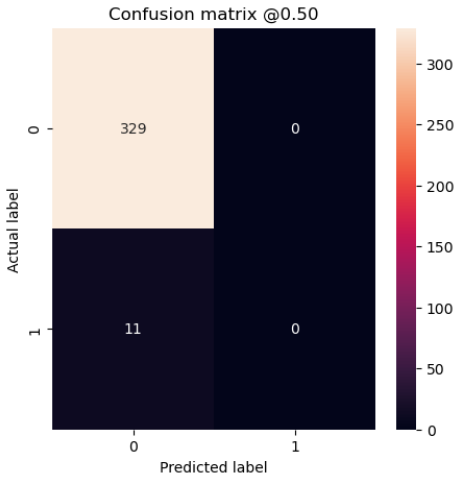

Figure 13 default MLP

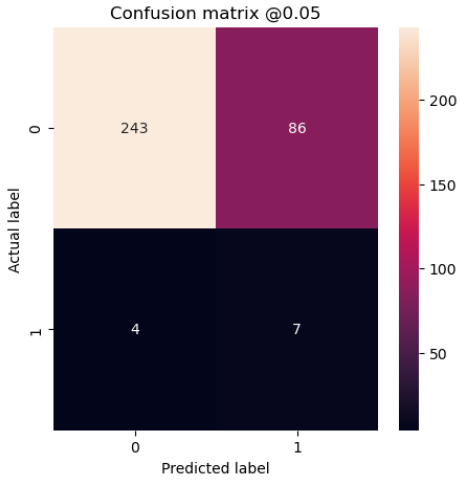

Figure 14 weighted MLP with threshold 0.05

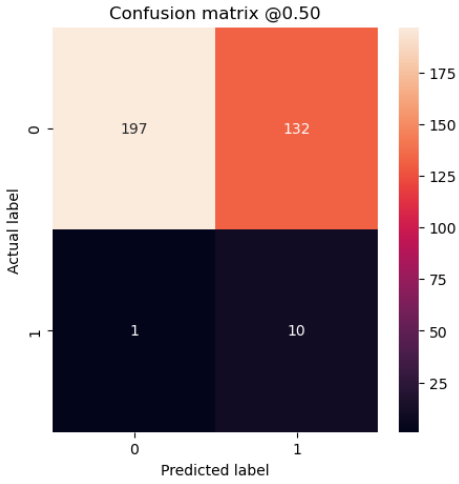

Figure 15 with Oversample

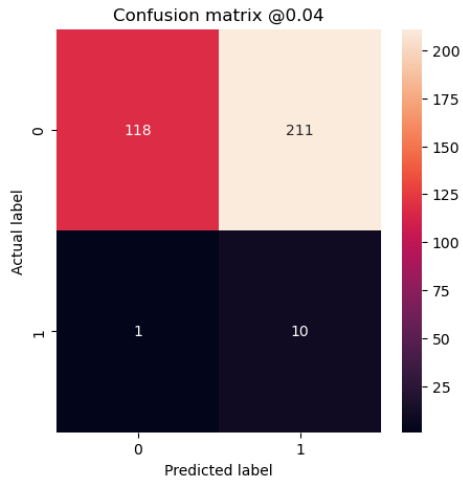

Figure 16 weighted MLP with threshold 0.04

Changes/Problems

In the original project plan, I planned to build 4 models base on 4 different time intervals: the time of admission to hospital, the end of the first day (24 hours after admission to the hospital), the end of the second day (48 hours after admission to the hospital), the end of the third day (72 hours after admission to the hospital), after feature selection the model random forest is based on top 5 features, these features are not excluded in any time intervals, in the end only 1 model is necessary for this project.

In the aim 2, Compare with the original project plan, I added the SVM model to compare with other models, in the MLP model training process, I used several techniques to improve the model performance and plotted several graphs to evaluate the performance of the model on different techniques, and tried to find the balance between precision and recall, but random forest performance is overall better than MLP.

[1]This template was adapted from the standard NSF research progress report (RPPR).

Problems or Delays Experienced and Corrective Actions Taken

One of the major problems experienced was that the model performance for the precision score always 0 for all models. To address this issue, I carried out an exploratory observation of the data distribution, I found that the dataset was too imbalanced, for this reason I applied a variety of techniques for rebalancing, such as oversampling, change class weight, change thredholds, undersampling. for rebalancing, each of these has a different degree of enhancement.

Impact

Impact on the Domain

Myocardial infarction (MI), colloquially known as "heart attack," is caused by decreased or complete cessation of blood flow to a portion of the myocardium. Myocardial infarction may be"silent," and go undetected, or it could be a catastrophic event leading to hemodynamic deterioration and sudden death. Most myocardial infarctions are due to underlying coronary artery disease, the leading cause of death in the United States. With coronary artery occlusion, the myocardium is deprived of oxygen. Prolonged deprivation of oxygen supply to the myocardium can lead to myocardial cell death and necrosis. Patients can present with chest discomfort or pressure that can radiate to the neck, jaw, shoulder, or arm. In addition to the history and physical exam, myocardial ischemia may be associated with ECG changes and elevated biochemical markers such as cardiac troponins. Through this project, the causes of the disease are identified, so that early detection and prevention can minimize the mortality rate.

Impact on the Individual

This project strengthened my coding skills and data science domain knowledge and put into practice what I learned in class, I found that metrics do not intuitively reflect the performance of the model, accuracy as a most commonly used metrics for imbalanced dataset is impractical, near perfect accuracy metric for imbalance dataset usually means that the model performance is poor, and cannot distinguish between negative and positive cases well. In practice, the model correctly videos a few cases that are of great significance, for example, suspicious transactions in bank records, probability of suffering from a major disease, and a few cases that are of great significance in these severely imbalanced data.

References:

"Myocardial Infarction." PubMed Central, National Center for Biotechnology Information, n.d. Myocardial Infarction - StatPearls - NCBI Bookshelf (nih.gov)

[1]This template was adapted from the standard NSF research progress report (RPPR).