Predicting Pulmonary Edema and Relapse for Myocardial Infarction Patients

Roshan Salil Parikh

Data Science Institute, Brown University

Providence, Rhode Island, United States of America

December 15, 2024

Updated December 21, 2024

Contents

1 Introduction

1.1 Purpose

Myocardial infarctions (MI), also known as heart attacks, affect over 1 million Americans every year. Two serious complication of MI include pulmonary edema, where fluid fills the lungs, and relapse of the MI. Pulmonary edema can develop suddenly after MI, and acute pulmonary edema requires immediate treatment. Relapse can occur suddenly or gradually, but still requires urgent medical attention. Machine learning prediction methods may allow hospital staff to be better prepared for potential complications for hospitalized MI patients [1].

1.2 Data

Data came from the UCI Myocardial Infarction Complications Dataset [2]. The data contains clinical information collected during each patient's hospital stay after suffering a MI and was collected over a 3-day period by Krasnoyarsk Interdistrict Clinical Hospital (Russia) between 1992-1995. The original dataset has 1700 examples and 111 features. There are 12 target variables available in the dataset, of which 'OTEK_LANC', which stood for pulmonary edema, and 'REC_IM', which stood for relapse, were used. For each target variable, a value of 1 indicated a diagnosis for that complication and a value of 0 indicated no diagnosis for that complication. Every example had missing values, but none of the target values did.

1.3 Previous Work

Previous work on this dataset conducted by researchers at Urmia University in Iran showed strong performance when predicting fatality from MI complications [3]. However, while f_1 scores were high, accuracy was close to or below baseline.

Newaz, et al., accounted for this dataset's imbalance through synthetic minority oversampling (SMOTE) and cost-based learning [4]. They predicted each of the target variables with different machine learning methods, obtaining accuracies close to baseline but relatively high precision.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to better understand the dataset with respect to the target variables. The data was imbalanced, with 9.40% of patients developing pulmonary edema and 9.40% of patients developing MI relapse. 2.36% of patients developed both complications.

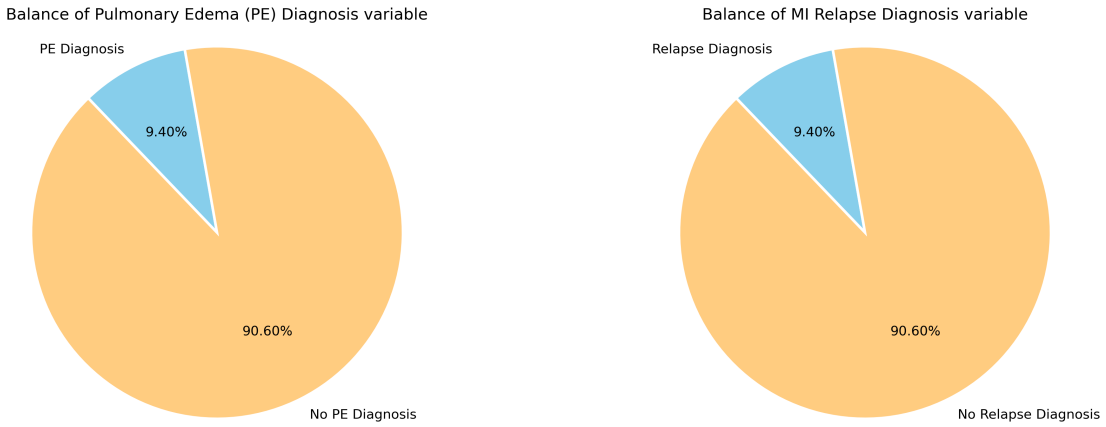

Figure 1: *Pie charts showing shares of pulmonary edema and relapse diagnoses.*

We first look at metrics available from bloodwork. Aside from MI, causes of pulmonary edema can include infection. White blood cell count and α-1 antitrypsin levels can increase during infection. However, EDA did not indicate a correlation between either and the target.

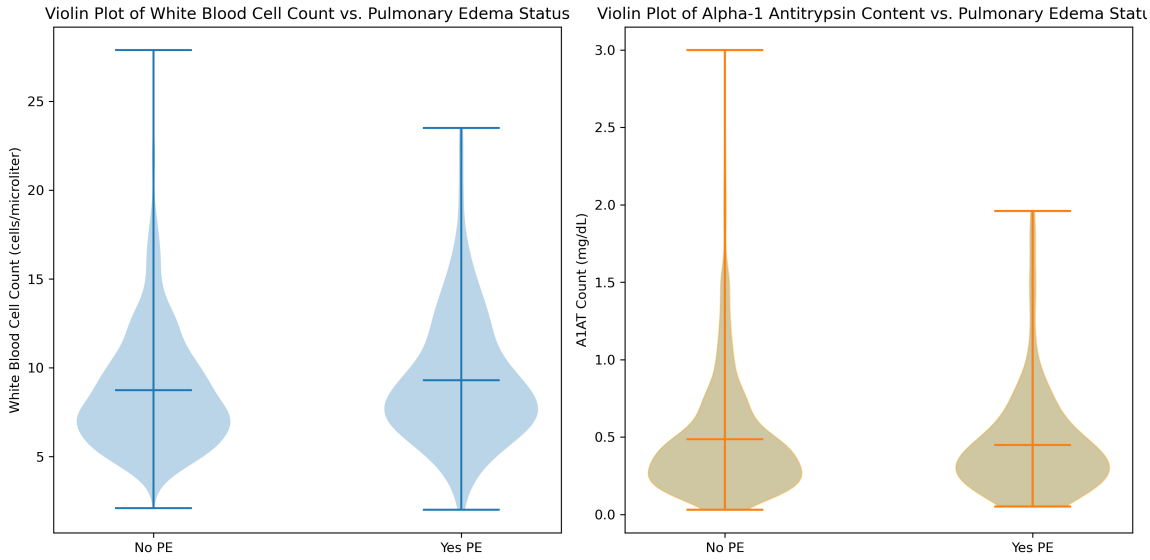

Figure 2: *Violin plots showing relationships between white blood cell count or α-1 antitrypsin levels and pulmonary edema.*

We also observe at the likelihood of each complication based on EKG readings.

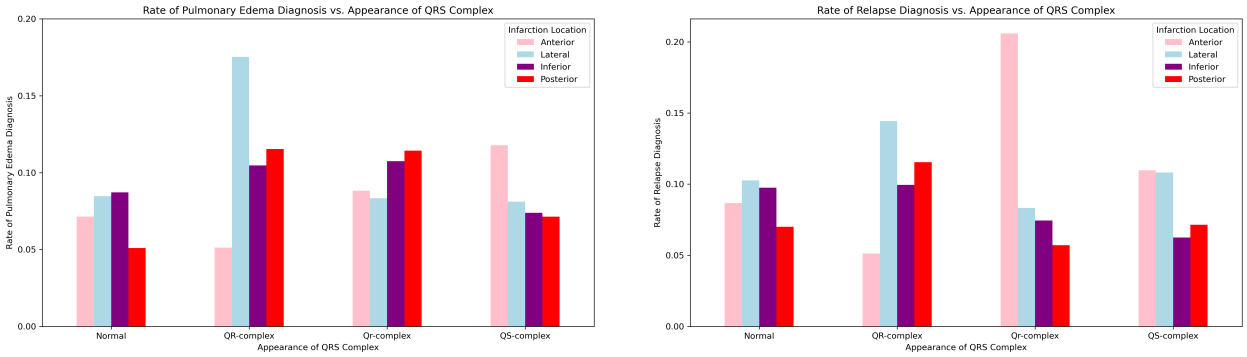

Figure 3: *The prevalence of pulmonary edema (left) and relapse (right) based on appearance of the QRS complex combined with the location(s) of the MI. A patient in this dataset could have MI in multiple locations, in which case their data would be included in multiple bars in this plot.*

However, EDA reveals that a patient is more likely to develop pulmonary edema if they are given liquid nitrates in the ICU. While included in the analysis, this variable may be misleading, since liquid nitrates can be given as a treatment for MI *and* as a treatment for pulmonary edema. The dataset did not specify when liquid nitrates were administered in relation to health events.

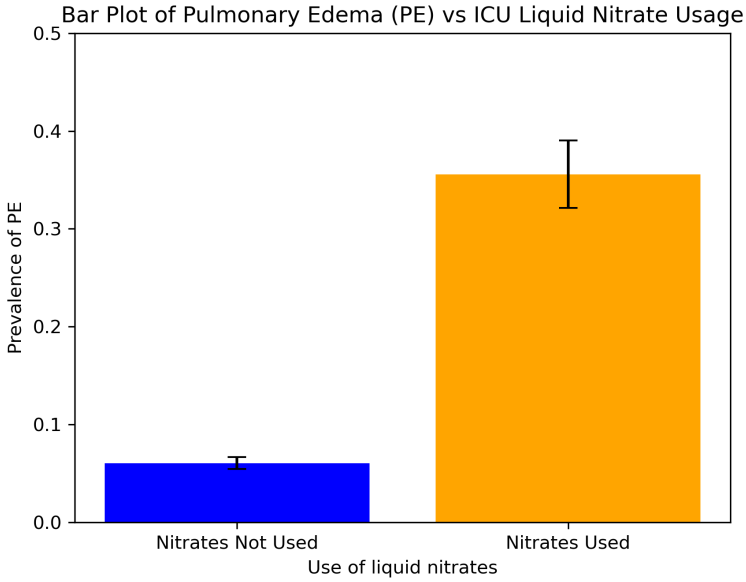

Figure 4: *Bar plots showing prevalence of PE diagnoses grouped by liquid nitrate usage in the hospital.*

Occurrence of chest pain on the third day of hospitalization appeared to show a pattern in relation to MI relapse.

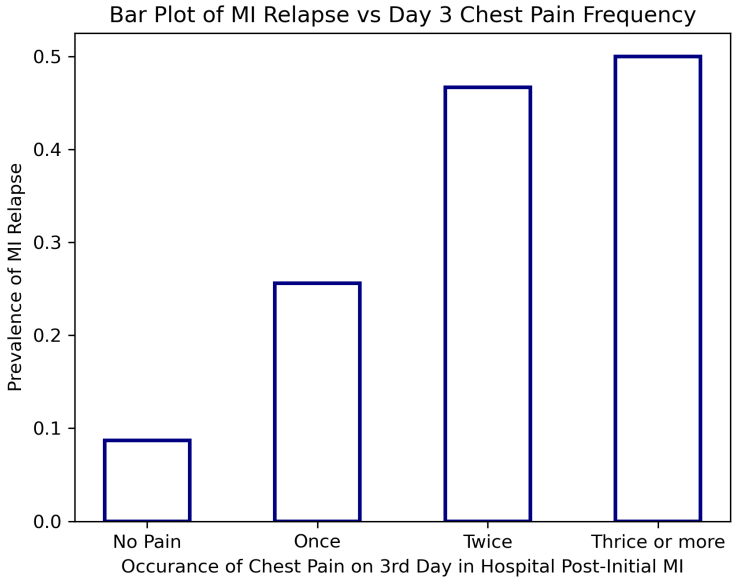

Figure 5: *Bar plots showing prevalence of relapse diagnoses grouped by occurrence of chest pain during the third day of hospital stay.*

3 Methods

3.1 Feature Selection

Features that had high Pearson correlations seemed to be clinically relevant, so no features were dropped due to this metric. Of the 111 features, only the 'SEX' feature did not have missing values. Examples that were missing age values were removed from the data, leaving 1692 examples. The 4 features (serum CPK content, heredity on coronary heart disease, and systolic and diastolic blood pressure in the ED) with more than 50% missing values were removed as well.

3.2 Splitting and Preprocessing

We first set aside 20% of the data to serve as a test set for our model. Because of the dataset's imbalance, we use scikit-learn's *StratifiedKFold* method with $n\_splits = 3$ to split the remaining data into training and validation sets, each with 902 examples. *StratifiedKFold* ensures an approximately equal distribution of classes between each of the training and validation sets.

Missing values in categorical and ordinal features were labeled as a new category, 'missing,' while iterative imputing was done on continuous features with missing features [1]. With scikit-learn's *ColumnTransformer*, blood pressure and age features are encoded with the *MinMaxScaler*, ordinal features with the *OrdinalEncoder*, and categorical features with the *OneHotEncoder*. All features were then standard scaled to have a mean of 0 and variance 1. The preprocessed data had 276 features.

[1] Iterative imputing was performed before implementing every model except for XGBoost, which can handle missing values in continuous features.

3.3 Evaluation Metric

f_1-score was used as evaluation metric during cross validation. Accuracy (ℓ_{0-1} loss) was also calculated and compared. Baseline scores were calculated by predicting the majority class (class 0) for accuracy and the minority class (class 1) for f_1-score over 5 tests sets, each with a different random seed.

3.4 Models Trained

Multiple types ($n = 5$) of classification models were tuned, trained, and tested for each target variable. Hyperparameter tuning was performed through cross-validation with the previously-mentioned Stratified K-Folds method and with scikit-learn's *GridSearchCV*. For each model, hyperparameters were tuned across different random seeds ($n = 5$) before the model with the highest f_1-score for each random seed was evaluated on the testing set. This process produced 25 trained models per target variable (50 trained models in total).

Method	**Hyperparameters tuned**
Logistic Regression (Ridge Regularization)	$C = \{10^n \mid n \in \{-8, \ldots, 3\}\}$ class weight $\in$ {balanced, none}
K-Nearest Neighbors	n neighbors $\in \{3, 5, 7, 10, 15, 30, 50, 70, 100\}$ weights $\in$ {uniform, distance} $p \in \{1, 2\}$
Support Vector Classifier with Linear Kernel	$C = \{10^n \mid n \in \{-5, \ldots, 3\}\}$ class weight $\in$ {balanced, none}
Support Vector Classifier with RBF Kernel	$C = \{10^n \mid n \in \{-5, \ldots, 3\}\}$ class weight $\in$ {balanced, none}
XGBoost	max depth $\in \{1, 3, 10, 30, 100\}$ column sample by tree $\in \{0.1, 0.25, 0.5, 0.75, 1\}$ positive class weighting $\in \{0.025, 0.05, 0.1, 0.25, 0.5, 1, 5, 10\}$

Table 1: *Classification methods and hyperparameters tuned through GridSearchCV.*

4 Results

4.1 Model Performance

We first observe results for when the target variable was pulmonary edema. On average, while models had high accuracies, they tended to be slightly lower than baseline of 91.15% $\pm$0.89 (by around 7.7%). f_1-scores were also generally low, with no single model's f_1 score surpassing 0.36. However, aside from K-Nearest Neighbors, f_1-scores were higher on-average than the baseline of 0.162 ± 0.015.

Results when Target = Pulmonary Edema

Method	**Avg. Accuracy - Baseline (%)**	**Avg. f_1-Score**
Logistic Regression	-10.4 ± 3.4	0.282 ± 0.064
K-Nearest Neighbors	-1.36 ± 0.92	0.0745 ± 0.063
Support Vector Classifier (Linear)	-9.32 ± 3.6	0.241 ± 0.095
Support Vector Classifier (RBF)	-10.9 ± 8.6	0.265 ± 0.077
XGBoost	-6.34 ± 0.99	0.303 ± 0.017

Table 2: *Average accuracy minus baseline accuracy (91.15%) and average f_1-score for each method. Models were trained and tested with 5 different random seeds.*

Because of the dataset's imbalance, we use the f_1-score as a measure of performance. The single best model was the SVC model with the RBF kernel ($f_1 = 0.358$), with parameters $C = 1$, class weight = balanced. On average, XGBoost performed the best ($f_1 = 0.303 \pm 0.017$). The single XGBoost model with the highest score ($f_1 = 0.319$) had the following parameters: column sample by tree = 0.25, max depth = 1, and positive class weighting = 5.

We now look at results for when the target variable is relapse of myocardial infarction. Average accuracies were much lower than the baseline of 91.45 ± 1.41 (by around 19.6%) than when the target variable was pulmonary edema, but f_1-scores followed a similar pattern to the previous target variable. The baseline f_1 score was 0.157 ± 0.024. All model types except for K-Nearest Neighbors outperformed the baseline f_1-score.

Results when Target = Relapse

Method	Avg. Accuracy - Basline (%)	Avg. f_1-Score
Logistic Regression	-22.4 ± 1.6	0.226 ± 0.019
K-Nearest Neighbors	-1.54 ± 1.20	0.0233 ± 0.0288
Support Vector Classifier (Linear)	-22.8 ± 2.1	0.212 ± 0.026
Support Vector Classifier (RBF)	-15.5 ± 4.3	0.228 ± 0.055
XGBoost	-16.2 ± 7.4	0.246 ± 0.036

Table 3: *Average accuracy minus baseline accuracy (91.45%) and average f_1-score for each method. Models were trained and tested with 5 different random seeds.*

We again use f_1 score as the metric for performance. The single best model was the SVC model with the RBF kernel ($f_1 = 0.333$) with parameters $C = 1$, class weight = balanced. On average, XGBoost had the best f_1-score of 0.246 ± 0.036, and the single XGBoost model with the highest f_1-score ($f_1 = 0.303$) had parameters column sample by tree = 0.1, max depth = 1, and positive class weighting = 5.

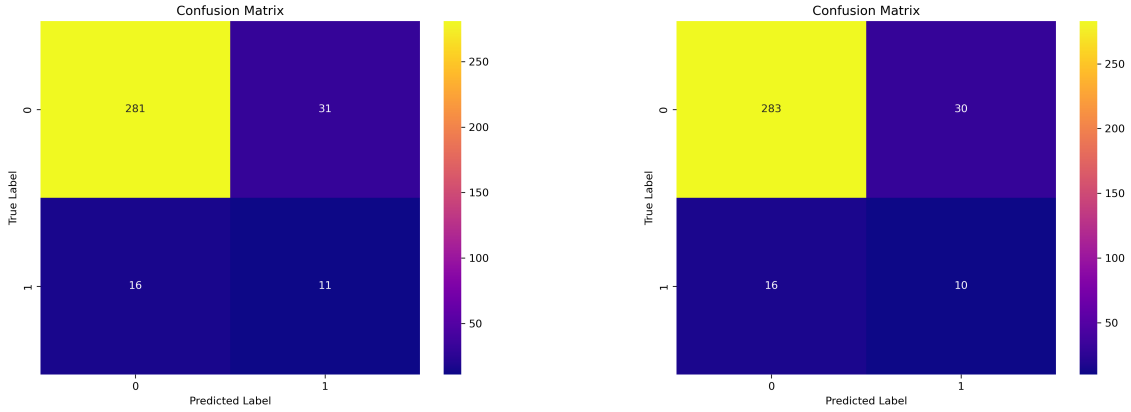

Figure 6: *Confusion matrices for the best XGBoost models when the target variable is pulmonary edema (left) and myocardial infarction relapse (right).*

For both targets, feature importance was assessed with the best performing XGBoost model, since XGBoost had the highest average f_1-score.

4.2 Global Feature Importance

Global feature importance was measured through various metrics, such as XGBoost's *gain* function, which measures a feature's contribution to reducing the log loss function when splitting a decision tree. Permutation importance was also calculated for each feature. Finally, SHAP values were calculated.

4.2.1 Pulmonary Edema Global Feature Importance

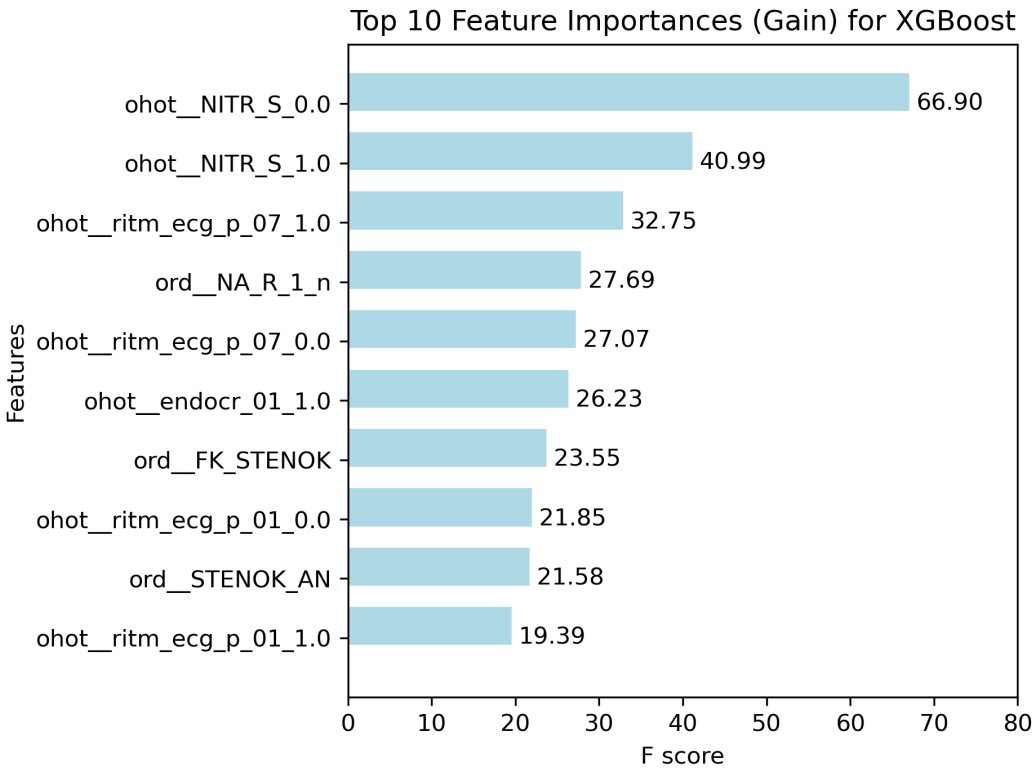

Figure 7: *Gain metric for features from XGBoost model with best f_1-score*

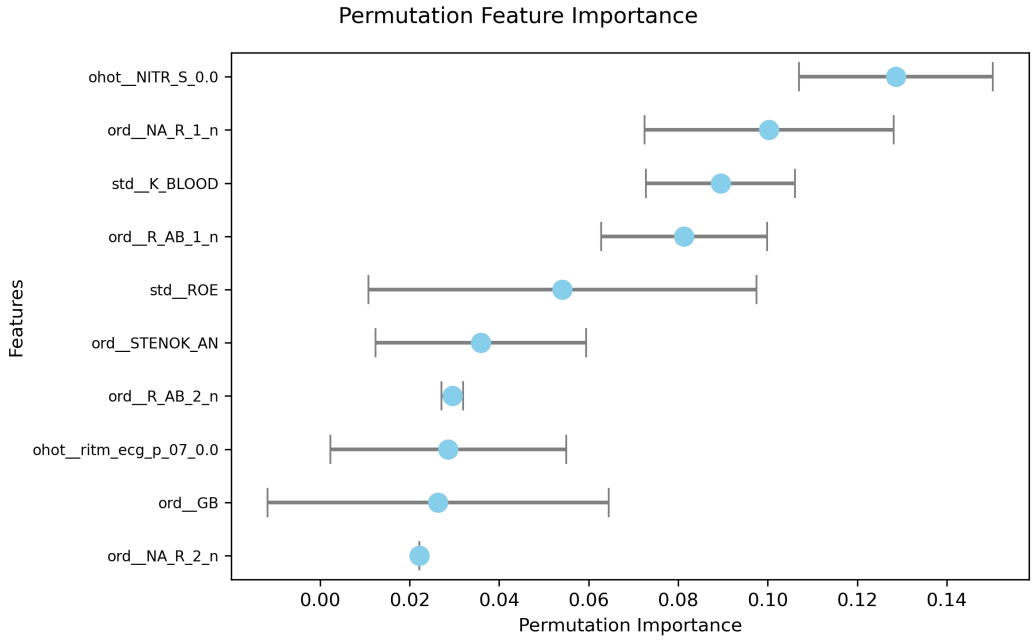

Figure 8: *Permutation importance values for features from XGBoost model with best f_1 score. f_1-score was used as the evaluation metric.*

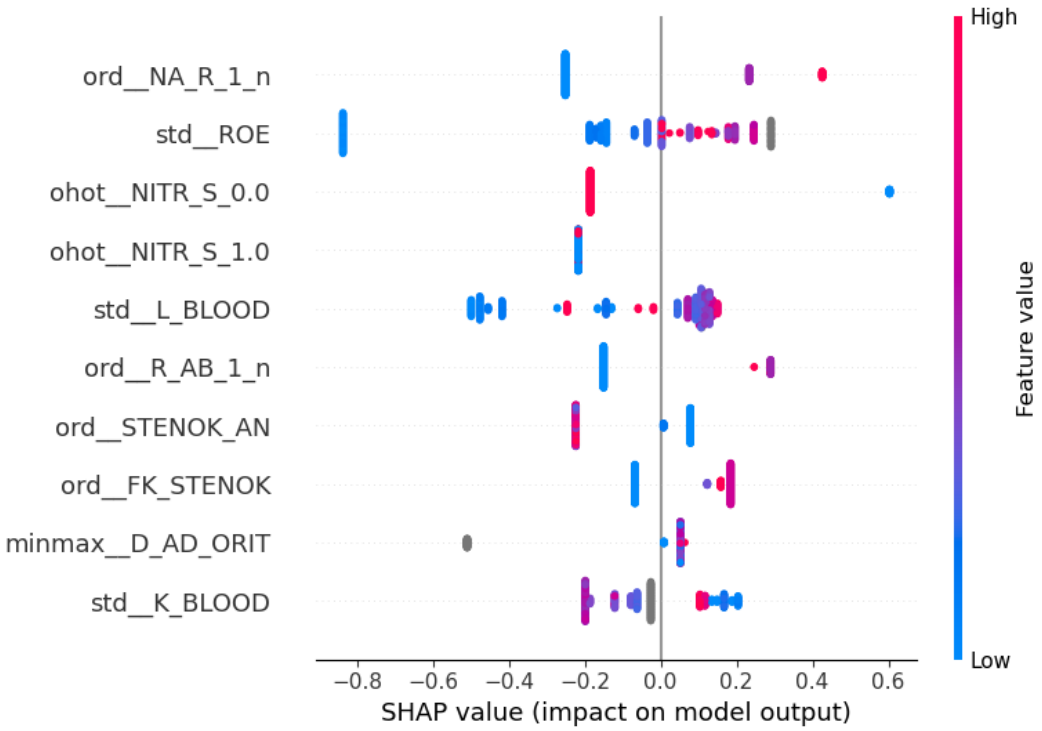

Figure 9: *SHAP values for features from XGBoost model with best f_1-score. f_1-score was used as the evaluation metric.*

Global feature importance analyses demonstrated that the administration of liquid nitrates ('ohot_NITR_S_0.0', 'ohot_NITR_S_1.0') was among the most important features for the model's predictive power, which followed patterns observed from EDA. Features corresponding to serum potassium levels ('std_K_BLOOD'), hospital opiod administration ('ord_NA_R_1_n', 'ord_NA_R_2_n'), EKG readings ('ohot_ritm_ecg_p_07_1.0', 'ohot_ritm_ecg_p_07_0.0'), and angina pectoris ('ord_FK_STENOK','ord_STENOK_AN') were also among the most important features.

4.2.2 Relapse of Myocardial Infarction Global Feature Importance

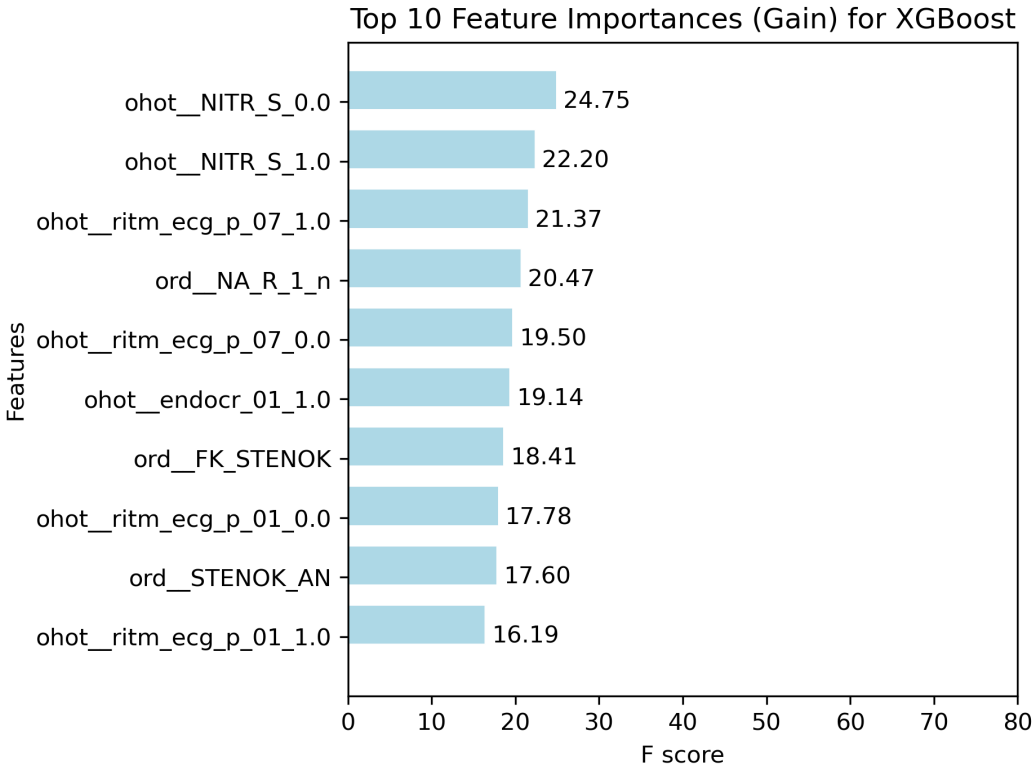

Figure 10: *Gain metric for features from XGBoost model with best f_1-score*

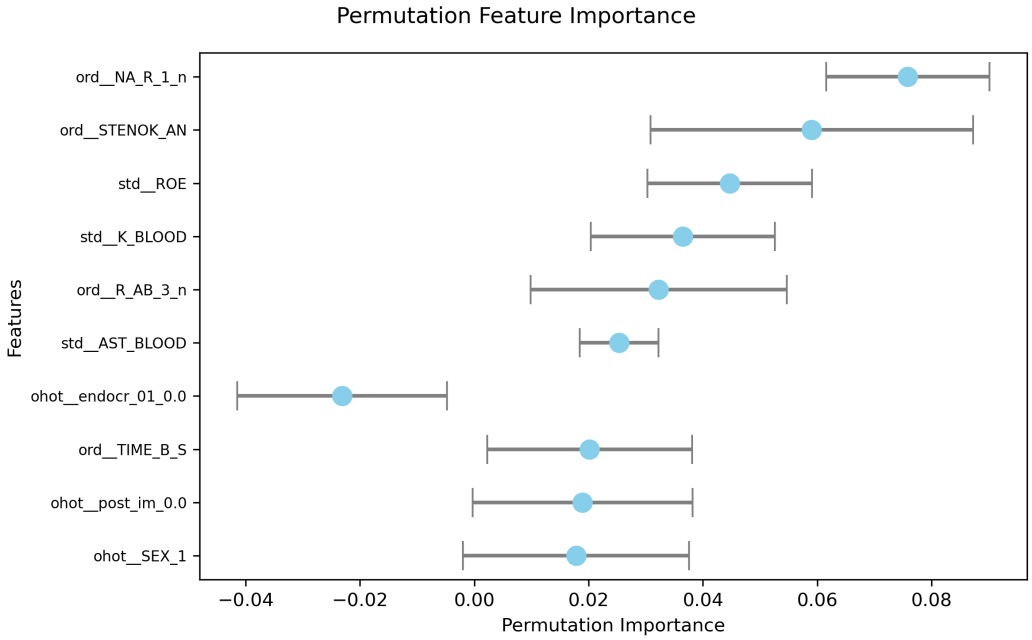

Figure 11: *Permutation importance values for features from XGBoost model with best f_1 score. f_1-score was used as the evaluation metric.*

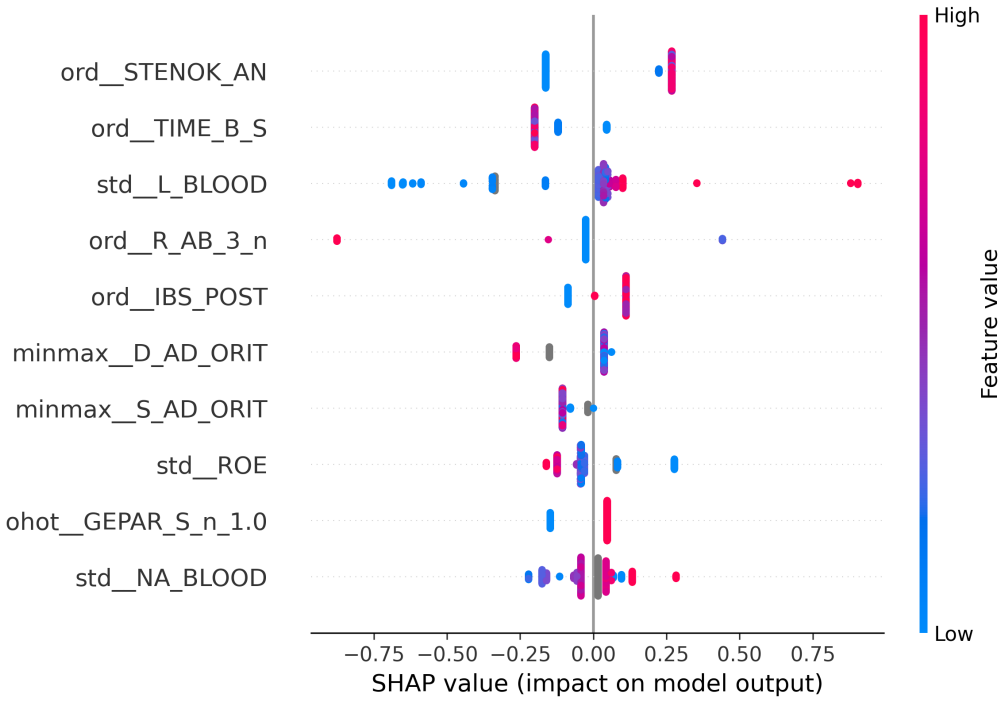

Figure 12: *SHAP values for features from XGBoost model with best f_1-score. f_1-score was used as the evaluation metric.*

Global feature importance analyses showed that features corresponding to the use of opioids during the hospital stay ('ord_NA_R_1_n') and angina pectoris ('ord_FK_STENOK', 'ord_STENOK_AN') tended to be among the most important features. Use of nitrates was considered relatively important through the gain method but not through other methods. Gain values were also relative low for the XGBoost model without the same variation seen for the other target variable, but SHAP analyses showed interesting separability of the features. For example, SHAP values indicate that history of external angina pectoris in the anamnesis had a positive correlation with relapse of myocardial infarction. Use of heparin in the ICU ('ohot_GEPAR_S_n_1.0') also positively correlated with relapse.

4.3 Local Feature Importance

SHAP values were calculated for specific data points to observe local feature importance. For each target variable, two points were arbitrarily chosen, of which one belonged to class 1.

4.3.1 Pulmonary Edema Local Feature Importance

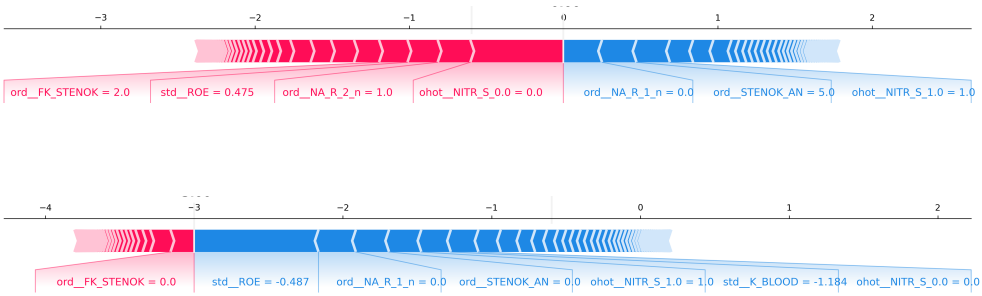

Figure 13: *SHAP force plots for points 2 (top, class 0) and 146 (bottom, class 1)*

.

One surprising observation is that erythrocyte sedimantation rate (ESR) ('std_ROE') was the most important factor for point 146. While ESR is not known to be directly correlated with

pulmonary edema, high ESR combined with preexisting cardiac conditions can increase the risk of developing pulmonary edema. Nitrate usage ('ohot_NIRT_S_0.0', 'ohot_NIRT_S_1.0') was also important for both data points, as were angina pectoris features.

Unsurprisingly, features associated with cardiac measurements and cardiac health history tended to be among the top 10 most important features.

4.3.2 Relapse of Myocardial Infarction Local Feature Importance

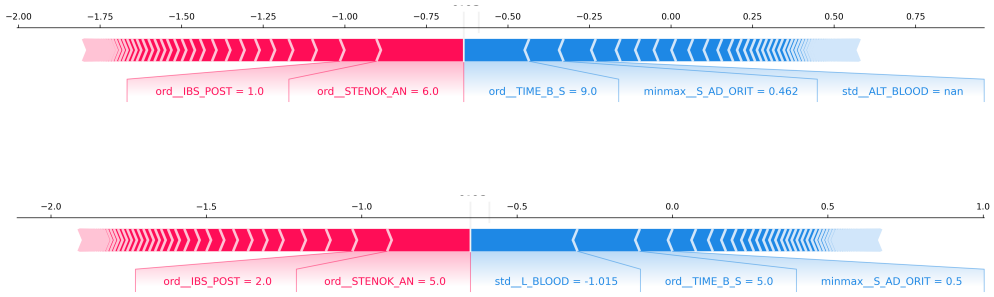

Figure 14: *SHAP force plots for points 0 (top, class 0) and 337 (bottom, class 1)*

For both data points, erythrocyte sedimentation rate ('std_ROE') was important when classifying the data point. In line with global SHAP feature analyses, angina pectoris history influenced the algorithm to predict the correct class for both points. Interestingly, white blood cell count ('std_L_BLOOD') was the most impactful feature when classifying point 337.

5 Outlook

Many factors may have led to the low performances of these model. First, the data contained clinical data collected over 3 days. Many of the features had metadata indicating when the data was collected, but many of them also did not or had imprecise time metadata. However,

while evaluation metrics were relatively low, feature analyses revealed interesting patterns about which features were most impactful to correctly classifying the data points.

When predicting pulmonary edema, nitrate administration appeared to be the most important factor by some metrics, but the reason for liquid nitrate usage was not provided, as they can be given as a response to either MI (therefore being a potential cause for pulmonary edema) or as a response to pulmonary edema. This complicates practical interpretation of the model.

SHAP values showed interesting patterns. For both targets, history of angina pectoris were considered important, more so for predicting relapse than for predictive pulmonary edema. SHAP values also showed that information learned from bloodwork, such as white blood cell count, sodium and potassium levels, and erythrocyte sedimentation rate, were useful in predicting both target variables.

Improving this model could require reducing the number of features and including the time when clinical measurements occured. This process may result in non-i.i.d. data, where each patient has multiple data points, but it may also result in a more powerful model. Data collected more recently and in the United States would also be more practical to eventual deployment of a predictive model, as cardiac care today in American hospitals may be significantly different from cardiac care of the 1990s in Russian hospitals. Supervised machine learning methods such as polynomial support vector classifiers may also be useful. Further, additional evaluation metrics, such as false positive rate, f_β score ($\beta \neq 1$), accuracy, and AUC, could provide more insights. Previous work indicates that fatality due to complications can be successfully predicted. The data may be more separable based on severity (or lethality) of MI complications as opposed to the presence of complications.

6 GitHub

Data, Jupyter Notebooks, and figures can be found at:

https://github.com/roshanparikh/HeartAttackComplications.git.

References

[1] M. Clinic. "Pulmonary edema." (2022), [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pulmonary-edema/symptoms-causes/syc-20377009. (accessed: 12.02.2024).

[2] S. Golovenkin, V. Shulman, D. Rossiev, *et al.* "Myocardial infarction complications." DOI: https://doi.org/10.24432/C53P5M. (2020), [Online]. Available: https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications.

[3] R. Ghafari, A. S. Azar, A. Ghafari, *et al.*, "Prediction of the fatal acute complications of myocardial infarction via machine learning algorithms," *The Journal of Tehran Heart Center*, vol. 18, no. 4, pp. 278–287, 2023. DOI: 10.18502/jthc.v18i4.14827. [Online]. Available: https://doi.org/10.18502/jthc.v18i4.14827.

[4] A. Newaz, M. S. Mohosheu, and M. A. Al Noman, "Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques," *Informatics in Medicine Unlocked*, vol. 42, p. 101 361, 2023, ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2023.101361. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823002071.