Watchout! Blood Clots Have Attacked Arteries in Russia: A study of Myocardial Infarction Complications at Anamnesis

Lennox Garay
MATH 533 - Statistical Learning
Cal State Fullerton
Fall 2023

Introduction

Machine Learning has the potential to revolutionize the healthcare industry. In many places, it already has. However, the American healthcare system is slow to adapt. Machine learning from patient demographic data is easily applicable even with current electronic medical records (EMRs). The standardization of EMRs and other diagnostic records may enhance patient care and reduce practitioner error. The use of ICD-10 codes are notoriously inconsistent metrics for diagnosis. Studies of the past have shown that there is little evidence to suggest that machine learning algorithms can learn from ICD-10 codes to predict diagnoses (Lingren et. al, 2016). It would be more useful to standardize the use of the EMR system to onboard various empirical patient data. The motivation of this paper is to demonstrate the confidence we should have in machine learning methods applied to medical data. Using different criteria, such as atrial fibrillation at anamnesis, could allow models to actively predict conditions in patients. Essentially, combining EMRs and specific patient data could improve the patient diagnostic process. The goal of this paper is to suggest that using such empirical patient data on various machine learning algorithms can be useful diagnostic tools with their high classification accuracy. Specifically, this paper implements many popular classification algorithms such as XGboost, logistic regression, support vector machine, decision trees, and basic multilayer neural networks. Such statistical modeling should be looked at as a tool for medical practitioners to boost confidence in patient care and reduce diagnostic errors. We will investigate the efficacy of computer models and their ability to classify patient diagnostic criteria specifically with regards to myocardial infarction complications.

Myocardial infarctions, or heart attacks, affect roughly 800,000 Americans every year. About 200,000 of these heart attacks occurred in individuals with at least 1 prior heart attack (NCHS, 2023). The biggest risk factor to heart attacks is heart disease, which affects 1 in 5 Americans. Globally, it affects nearly 200 million individuals. Meaning globally it only affects 2.5% of the population. Comparatively, It's clear that America has a problem. The issue regarding this project refers to the complications that may arise from myocardial infarctions. It's quite likely that individuals who have one heart attack are prone to having another. There are worse complications beyond this. Figure 1 is a list of the complications studied and observed.

This project seeks to provide evidence for building confidence in computational methods in medical diagnoses. We begin with the data at hand. The data is from a hospital in Krasnoyarsk, Russia. There were 1700 observations of patients who presented to the hospital with acute myocardial infarction. There were 111 predictors and 12 response variables. Some of the predictors include empirical data such as systolic and diastolic blood pressure. The majority of predictors were binary categorical values. Some examples include diabetes, history of stroke, tobacco use, and ECG readings. The complications are seen below.

Complication
Atrial fibrillation (FIBR_PREDS)
Supraventricular tachycardia (PREDS_TAH)
Ventricular tachycardia (JELUD_TAH)
Ventricular fibrillation (FIBR_JELUD)
Third-degree AV block (A_V_BLOK)
Pulmonary edema (OTEK_LANC)
Myocardial rupture (RAZRIV)
Dressler syndrome (DRESSLER)
Chronic heart failure (ZSN)
Relapse of the myocardial infarction (REC_IM)
Post-infarction angina (P_IM_STEN)
Lethal outcome (cause) (LET_IS)

(Figure 1.)

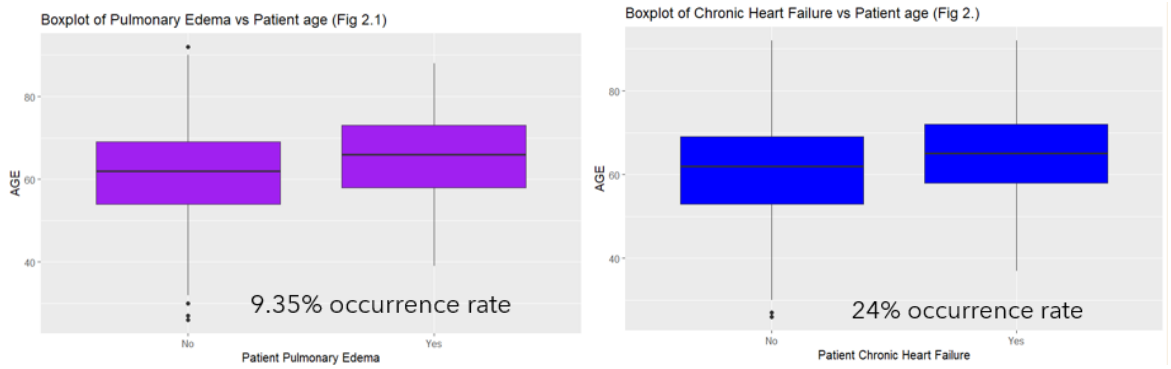

Figure 2.

Fibrillation and Tachycardia are essentially the heart beating out of baseline whether it's arrhythmic or *Affrettando*. Atrial fibrillation specifically refers to the unusual/arhythmic beating. Tachycardia refers to beating faster than normal. Third-degree AV BLOCK is a serious condition. It occurs when there is a complete lack of communication between atria. The electrical impulses from the top of the heart do not reach the bottom part of the heart. This can prevent blood from reaching the heart and brain. Dressler syndrome is inflammation of the sac surrounding the heart (pericarditis). It's an immune response to myocardial damage as a result from heart attack, surgery, or other traumatic injury. Figure 2 shows two of the twelve complications explored in this project. These two complications (pulmonary edema and chronic heart failure) had the highest incidence rate of 9.35% and 24%, respectively. We will use various machine learning methods to predict the complications in figure 1 using the covariates listed in the variable description page[1].

Methods

It is difficult to ascertain a complete list of valid covariates for this predictive classification task. This project used the covariates listed in the data documentation (Golovenkin, 1995). The models used in this classification analysis were: logistic

regression, XGboostclassifier, KNN, linear support vector machine, radial basis function support vector machine, multilayer perceptron, and decision trees. There was a 80/20 split for the training and testing data across all these models. The python code for generating the models can be found in the appendix. All of the response variables in the data were binary categorical variables excluding the LET_IS variable, which indicates living status (0) or cause of death (1-7). To simplify the computation process, the LET_IS variable was rewritten to be a binary variable for living and death status as 0 or 1, respectively. Nearly every patient recorded in this data set had one missing value. The data required imputation for most of these models. The imputation method replaced missing values with the median. The models were all fed the same parameters for training and validation data. The decision tree and XGBoost models were initiated using *max_depth = 3 and num_params = 1,* for the models, respectively. Below are some brief details of select models used in this presentation.

Decision Trees

A supervised machine learning algorithm that represents class labels as leaves. Branches represent groups of features that lead to said labels. This algorithm is essentially a recursive binary tree that maximizes information gain using entropy through the following equations:

$$H(T) = -\sum_{i=1} p_i log_2(p_i)$$

$$IG = H(T_{parent}) - \sum w_i H(T_{child_i})$$

Where p_i is the proportion of data points that belong to class k.

Extreme Gradient Boosting

Excellent at handling missing data. Like ridge regression in that we are minimizing the loss function which is the sum of square residuals plus a penalty or regularization term. The essence of XGB is that we are now taking that loss function and minimizing it using a 2nd Order Taylor approximation. Plug in residuals divided by previous probability plus regularization term to get the optimal output value. It uses this to compare similarity scores and make a decision at the nodes. The following formulas summarize this information:

$$[\sum_{i=1} L(y)i, p_i^0 + O_{value} + \frac{1}{2}\lambda O_{value}^2$$

$$L(y_i, p_i) = -[y_i \log(p_i) + (1-y)\log(1-p_i)]$$

$$O_{value} = -\frac{\sum \mathbf{g}}{\sum \mathbf{h} + \lambda}$$

$$\text{Similarity score} = \frac{\sum \mathbf{g^2}}{\sum \mathbf{h} + \lambda}$$

$$\text{Gain} = Sim_{left} + Sim_{right} - Sim_{root}$$

Where g and h are the gradient and hessian, respectively. Like in decision trees, p is the proportion of data belonging to true class at node 0.

Support Vector Machine

Classification algorithm used to differentiate class belonging in linearly separable data. SVM minimizes the following soft margin to classify data

$$[\frac{1}{n}\sum_{i=1} max(0, 1 - y_i(w^T x_i - b))] + \lambda||w^2||$$

[1]Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer

If that data is non-linear we can use the radial basis function kernel to transform our

	FIBR_PREDS	PREDS_TAH	JELUD_TAH	FIBR_JELUD	A_V_BLOK	OTEK_LANC	RAZRIV	DRESSLER	ZSN	REC_IM	P_IM_STEN	LET
LogisticRegression	85.29412	98.82353	96.76471	94.70588	95.29412	88.52941	96.17647	94.11765	75.58824	89.11765	87.64706	86.76
XGBoostClassifier	85.88235	99.41176	97.64706	95.58824	96.17647	89.41176	96.47059	93.82353	80.29412	89.70588	89.11765	93.52
KNN	84.70588	99.11765	97.64706	96.17647	96.76471	90.0	96.47059	93.82353	74.70588	90.29412	88.23529	84.41
Linear SVM	85.58824	99.11765	97.64706	96.17647	96.76471	89.70588	96.47059	93.82353	75.58824	90.58824	89.70588	87.64
RBF SVM	85.88235	99.41176	97.64706	96.17647	96.76471	90.29412	96.47059	93.82353	76.76471	90.88235	89.70588	83.82
DecisionTree	84.70588	98.52941	97.35294	95.58824	95.0	87.94118	93.23529	93.52941	79.41176	88.82353	88.52941	89.11

Table 1.

classification criterion. Instead of having a linear classifier we can use a kernel

$$k(x, y) = exp(-\frac{||x - y||^2}{2\sigma^2})$$

Where σ^2 is a free parameter. This radial basis function kernel allows us to classify the data at hand because it is largely non-linearly separable. The computation of the kernel function is the high (infinite in RBF) dimensional relationship between observations that are relatively close or far away from each other. Where the data are close to each other the value of k(.) is near zero.

Results and Discussion

A summary of the models and their performance for each of the response variables are shown in Table 1. The (i,k) element of Table 1 represents the accuracy of model i in predicting response variable k. Looking at Table 1, we see that FIBR_PREDS has roughly an 85% classification rate. There is not much variance between the models in classifying this response variable. This means that across models, the classification accuracy for atrial fibrillation is roughly 85% with the best model being XGBoost with 85.88% accuracy. This seems to be the trend with all of the columns in the accuracy table. Each columns in the accuracy table show that there is not too much variance between each of the models for predicting the response variable. This might suggest that the models do not specifically understand the intricacies of the data.

For example, all of the models for PREDS_TAH formally known as supraventricular tachycardia have a classification rate of 99%. This is a clear sign that there is something wrong with the model. Machine learning convention discourages faith in models that have 99% accuracy across the different models. The suspected reason for such high classification accuracy is not over-fit on the models specifically, but the data at hand. The issue with the supraventricular tachycardia data is

[1]Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer

that hardly any patients presented to the hospital with this condition. Considering that the incidence rate was less than 5%, the models are guessing that nobody has tachycardia, resulting in a high classification accuracy. There is more to be done on investigating supraventricular tachycardia specifically. In reality, each column in this data set should deserve more detailed treatment for producing medical-grade diagnostic tools. However, the significance here is to demonstrate their potential as such tools rather than create these tools within this paper. Future research on this will be discussed in the conclusion.

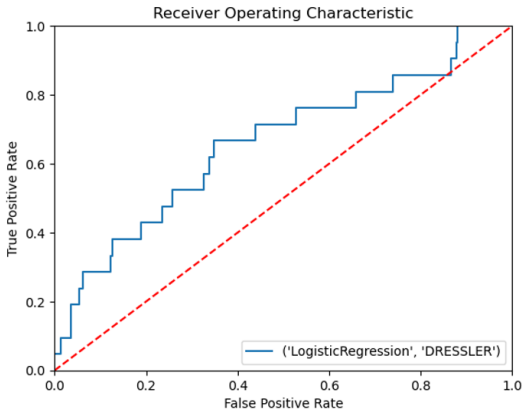

Figure 3.

Returning to the models at hand, logistic regression was only good at predicting Dressler syndrome (pericarditis). Figure 3 shows the ROC curve for the model predicting Dressler syndrome. Each model for each response variable has its own ROC plot, meaning there are 72 such plots. For the sake of brevity these plots will be included in the python code attached to the appendix.

It's interesting that the logistic regression model was the best at predicting Dressler's Syndrome. Dressler's syndrome is diagnosed using expensive testing. Although no criteria for the diagnosis of acute pericarditis have been established, prior studies have suggested that at least 2 of the following 4 criteria should be present: (1) characteristic chest pain, (2) pericardial friction rub, (3) suggestive electrocardiographic (ECG) changes, and (4) new or worsening pericardial effusion (Khandakher et. al, 2010). It may be worth investigating whether or not all tests are absolutely necessary for diagnostic criteria. That is where these machine learning algorithms may come in handy. If there is skepticism on the diagnostic criteria for certain diseases or conditions, then it may be worth investigating using computational methods. In this case, it may be worth using logistic regression to investigate the significance of various empirical patient data to evaluate the risk of having these diseases such that it reduces the need for 4 expensive medical tests.

There was an implementation of a neural network on this data, however, there was no significant classification accuracy using the multilayer perceptron. The model was invalid with this data. Specific model parameters included 3 layers, a batch size of 32, and an adam optimizer for binary classification. This may have to do with there being more than binary data in the dataset. A neural network in patient data may require more careful treatment in a

[1]Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer

specialized paper.

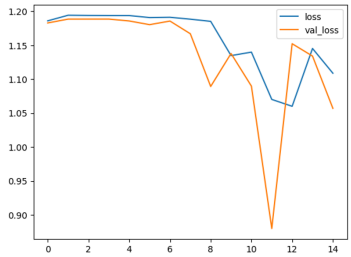

Figure 4.

We can see from the loss plot that the loss explodes, likely due to bad data for this type of model. For brevity, model diagnostics will not receive a full treatment, but some other issues present in this model could include data with high variance, or division by 0 in the data, which could be true sometimes despite standardization of data. Standardization of categorical data could be problematic for neural networks. For example, SVM and logistic regression do not heavily weigh the unscaled categorical variables. Another issue present is that much of the data is abstracted simply from the sheer volume. It's hard to know what data is where and how it's written. Again, for brevity, this is omitted in neural network implementation. This paper offers many more competitive models for this dataset such as XGboost and SVM.

Conclusion

This paper has discussed the efficacy of different machine learning models in the field of medicine. The two fields have been intimately involved for a long time, but there is still much work to be done. Specifically, it may be worth investigating the integration of EMRs and empirical health data such that machine learning methods can be implemented directly into the EMR or some interface of it. This way EMRs become an integrated diagnostic tool for medical practitioners. In a data set such as the one used in this paper, it's clear that categorical or empirical data work in tandem to provide highly accurate classification models. This should inspire confidence in building medical grade models where each column in our accuracy table receives its own special treatment and hyperparameters tuning for optimal results. The goal of this paper was to provide a basis for further research. The models in this paper are nowhere near production grade, but with highly collaborative work between statisticians and medical scientists, these models and EMR integration can become highly valuable diagnostic tools that reduce costs for patients in the long run. This type of computational integration will support a streamlined diagnostic process that provides clarity for both practitioner and patient.

It's important that medical diagnostic tools advance as quickly as the technology available to private companies. The significance here is that there are many categorical variables that do not require expensive screenings, which allow models like these to provide cheap and effective diagnostic criteria. It's imperative to consistently improve patient care quality and the integration of EMRs and empirical patient data are a step in the right direction.

[1]Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer

REFERENCES

Khandaker MH, Espinosa RE, Nishimura RA, Sinak LJ, Hayes SN, Melduni RM, Oh JK. Pericardial disease: diagnosis and management. Mayo Clin Proc. 2010 Jun;85(6):572-93. doi: 10.4065/mcp.2010.0046. PMID: 20511488; PMCID: PMC2878263..

Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, Bickel J, Wildenger Welchons L, Reinhold J, Bing N, Ni Y, Barbaresi W, Mentch F, Basford M, Denny J, Vazquez L, Perry C, Namjou B, Qiu H, Connolly J, Abrams D, Holm IA, Cobb BA, Lingren N, Solti I, Hakonarson H, Kohane IS, Harley J, Savova G. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. PLoS One. 2016 Jul 29;11(7):e0159621. doi: 10.1371/journal.pone.0159621. PMID: 27472449; PMCID: PMC4966969

Mechanic OJ, Gavin M, Grossman SA. Acute Myocardial Infarction. [Updated 2023 Sep 3]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-.
Myocardial infarction complications Database. (2020). Leicester.figshare.com. https://doi.org/10.25392/leicester.data.12045261.v3

Ojha, N., & Dhamoon, A. S. (2022). Myocardial Infarction. National Library of Medicine; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK537076/

Ornato JP, Hand MM. Warning signs of a heart attack. Circulation [J]. 2014 Mar 18; 129(11):e393-5.
https://www.cdc.gov/heartdisease/facts.htm, 2020

S.E. Golovenkin, Alexander GorbanAlexander Gorban, Evgeny MirkesEvgeny Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, M.G. Dorrer 2020

Steeg, G. V., & Galstyan, A. (2015, January 30). Maximally Informative Hierarchical Representations of High-Dimensional Data. ArXiv.org. https://doi.org/10.48550/arXiv.1410.7404

[1]Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer