Creating A Machine learning Model to Predict Presence of Coronary Artery Disease

Long Nguyen, Marek Boulerice, Sarah Eshafi, Hui Tang

Table of contents

1. SUMMARY

The following document covers a machine learning model analysis with a goal to predict angiographic coronary disease in patients. Data is pulled from patients undergoing angiography at the Cleveland Clinic in Ohio. This analysis is composed of Exploratory Data Analysis, testing of various machine models on a training data set, model optimization via hyperparameter, and final model performance analysis. The final model selected is a LogisticRegression Model, which performs quite well on unseen data. The model returns an F1 score of 0.808, while maintaining an accuracy of 0.828 on test data. Overall the model is quite promising in its potential applications. The high recall value of 0.7 is encouraging, signifying the model reports returns very few few missed cases of angiographic coronary diseases. Nevertheless, in a case as serious as predicting heart diseases even a few missed diagnoses are concerning. As such, we believe further tuning of the model should be performed. As well, there are some limitations to the model developed. The model requires significant amount of patient data to return an accurate prediction, which is collected in an in-clinic examination, at which point a diagnosis could likely be delivered regardless. Nevertheless, the results of the report show potential for a detection tool such as this to be used in the future, given further improvements.

2. INTRODUCTION

Heart disease is the leading cause of death worldwide (Organization 2024). Treating these heart diseases depends on capability of detecting symptoms and diagnosing cases earlier. One complication to diagnosing heart diseases is that many cases are found to be asymptomatic (Master 1969). This creates an opportunity for application of machine learning methods, where the following question can be asked: Given various details about a client's medical status, can we create a statistical model to accurately predict whether the patient has the disease? The goal of the following analysis is to create a model that can answer this question. To be best suited to the problem, the model should retain a high accuracy while minimizing the number of false negatives (i.e. predicting that a patient does not have the heart disease when the patient in fact does).

In particular, this analysis focuses on detection of angiographic coronary disease. The data set used in creating our model was taken from 303 patients undergoing angiography Cleveland Clinic in Cleveland, Ohio (Janosi and Detrano 1989). From this procedure, a set of parameters were collected on each patient, and a diagnosis of whether the patient had the angiographic coronary disease (signified by a diameter narrowing of the coronary artery by at least 50%). This is to serve as the target variable in our analysis.

The set of parameters collected during the procedure, used as our features for model training, are as follows:

- **Age (in years)** : Age of patient (years)
- **Sex** : Sex of patient (male or female)
- **Chest pain type**: categorical feature describing the type of pain experienced by the patient
- **Resting Blood Pressure**: numeric feature giving patients resting blood pressure
- **Serum Cholesterol** : numeric feature giving the patients Serum cholesterol in mg/dl
- **Fasting blood sugar > 120 mg/dl** : binary feature indicating whether the patients Blood sugar level while fasting exceeded 120 mg/dl
- **Resting electrocardiographic results**: categorical feature reporting patients ECG results
- **Maximum heart rate achieved**: numeric feature giving maximum heart rate achieved by patient
- **Exercise-induced angina**: binary feature indicating whether patient underwent exercise induced angina
- **ST depression induced by exercise relative to rest**: numeric feature indicating the ST depression induced by exercise relative to rest
- **Slope of the peak exercise ST segment**
- **Thalassemia** : categorical feature indicating if patient suffered from Thalassemia

The following sections will discuss the decisions made and results in our Exploratory Data analysis, Machine learning model training, and final model performance.

This report also drew information from the study done by (O'Flaherty et al. 2008) and (Athanasios Aessopos 2007)

3. DATA VALIDATION & CLEANING

From an initial preview of the data, some issues were corrected immediately, before performing formal data validation. This includes removing invalid target values (i.e. values outside of range) and converting the values to their semantic meaning. The output of the data cleaning can be seen in Table 1 below.

Table 1: Preview of cleaned data.

Age (in years)	Sex	Chest pain type	Resting blood pressure (in mm Hg on admission to the hospital)	Serum cholesterol (in mg/dl)	Fasting blood sugar > 120 mg/dl	Resting electrocardiographic results	Maximum heart rate achieved	Exercise induced angina	ST depression induced by exercise relative to rest	Slope of the peak exercise ST segment	Number of major vessels (0–3) colored by fluoroscopy	Thalassemia	Diagnosis of heart disease
55	male	atypical angina	130	262	False	normal	155	no	0	upsloping	0	normal	< 50% diameter narrowing
51	male	typical angina	125	213	False	showing probable or definite left ventricular hypertrophy by Estes' criteria	125	yes	1.4	upsloping	1	normal	< 50% diameter narrowing

Table 1: Preview of cleaned data.

Age (in years)	Sex	Chest pain type	Resting blood pressure (in mm Hg on admission to the hospital)	Serum cholesterol (in mg/dl)	Fasting blood sugar > 120 mg/dl	Resting electrocardiographic results	Maximum heart rate achieved	Exercise induced angina	ST depression induced by exercise relative to rest	Slope of the peak exercise ST segment	Number of major vessels (0–3) colored by fluoroscopy	Thalassemia	Diagnosis of heart disease
55	female	asymptomatic	128	205	False	having ST-T wave abnormality	130	yes	2	flat	1	reversable defect	>50% diameter narrowing
67	male	asymptomatic	120	237	False	normal	71	no	1	flat	0	normal	<50% diameter narrowing
51	female	non-anginal pain	140	308	False	showing probable or definite left ventricular hypertrophy by Estes' criteria	142	no	1.5	upsloping	1	normal	<50% diameter narrowing

Next, we evaluated whether the features in the train and test dataset are distributed similarly

using deepchecks. The check concluded that that the features were not distributed significantly and our distributions are as expected.

4. METHOD

The following section outlines the steps taken in manipulating our data and creating our model.

EDA is first conducted to obtain an idea of feature importance and to establish any important correlations to watch for. Machine learning analysis is then performed, where multiple models are tested and their performance compared. The best performing model is selected to proceed with. On this model, hyperparameter optimization is completed via random search to tune our model and obtain best results. Finally, the model is trained and tested on a separate data set and evaluated for performance.

4.1 EDA

In this section, preliminary analysis is conducted to obtain an idea of possible correlations between features to be on the look out for. The results are presented below.
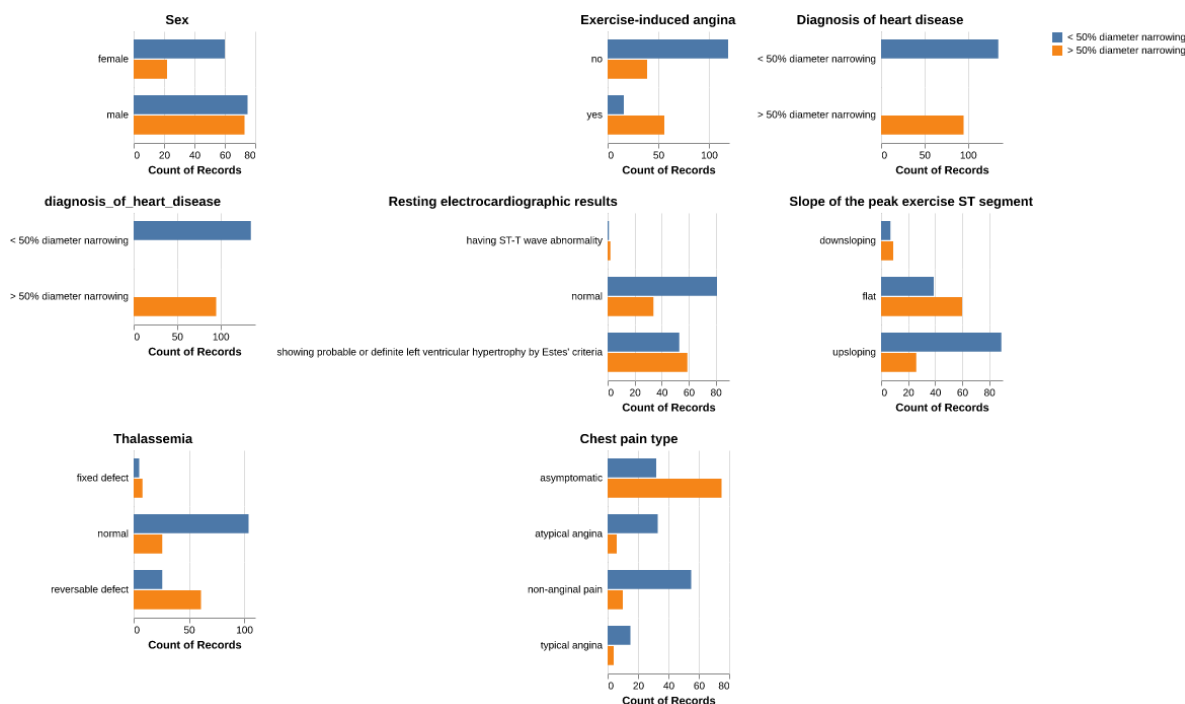

Figure 1: Distributions of Categorical Features

5

From Figure 1 above, we can see that there is some class imbalance in the target (diagnosis of heart disease). This means we will need to consider balanced scoring metrics and models.
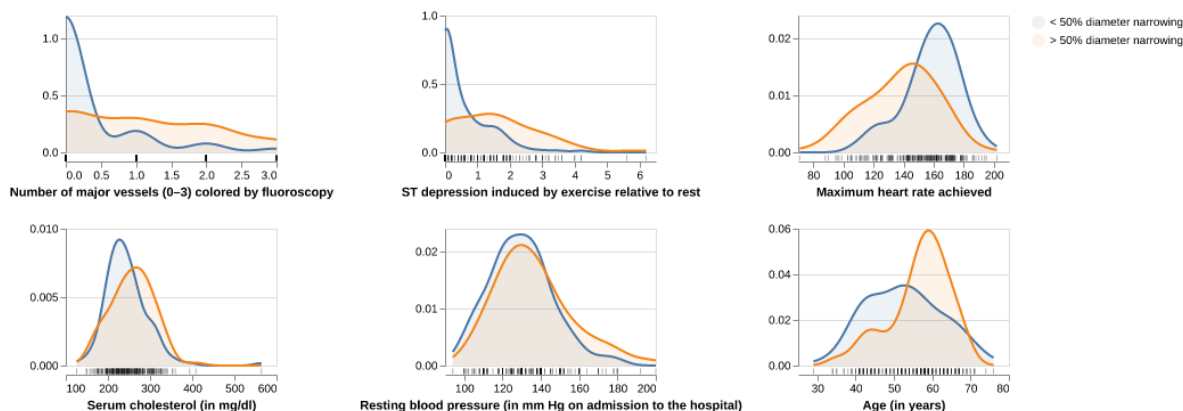

Figure 2: Distributions of Numeric Features

Figure 2 provides insights on some potential useful features for prediction. For example, it appears that patients testing negative for heart disease can achieve a higher maximum heart rate.

Finally, in Figure 3, we can see that there are correlations in some variables.

4.2 ML-ANALYSIS

The following section outlines the procedure taken in creating our model and testing it on our data set. As this is a classification problem (predict whether the patient has the disease or not), the chosen models for testing in this analysis are a Logistic Regression model and a Support Vector Classifier. These two models were selected as they have been shown to historically perform well on real world data sets.

Since this data set is somewhat imbalanced, the primary scoring metric to evaluate these models will be F1 score, though model accuracy is still taken into consideration. Due to the nature of the analysis, special attention is taken to minimize false negatives as they represent the most damaging type of error (predicting that a patient is free of the disease when he does in fact have it). To this effect, we look to maximize Recall.

The framework upon which this analysis is based has been adapted from DSCI573 Lab1.

4.2.1. DATA PREPROCESSING

Features are sorted by type, and a column transformer object is created. On categorical columns, simple imputing is applied filling missing values with the most frequently occurring value. One hot encoding is then performed.
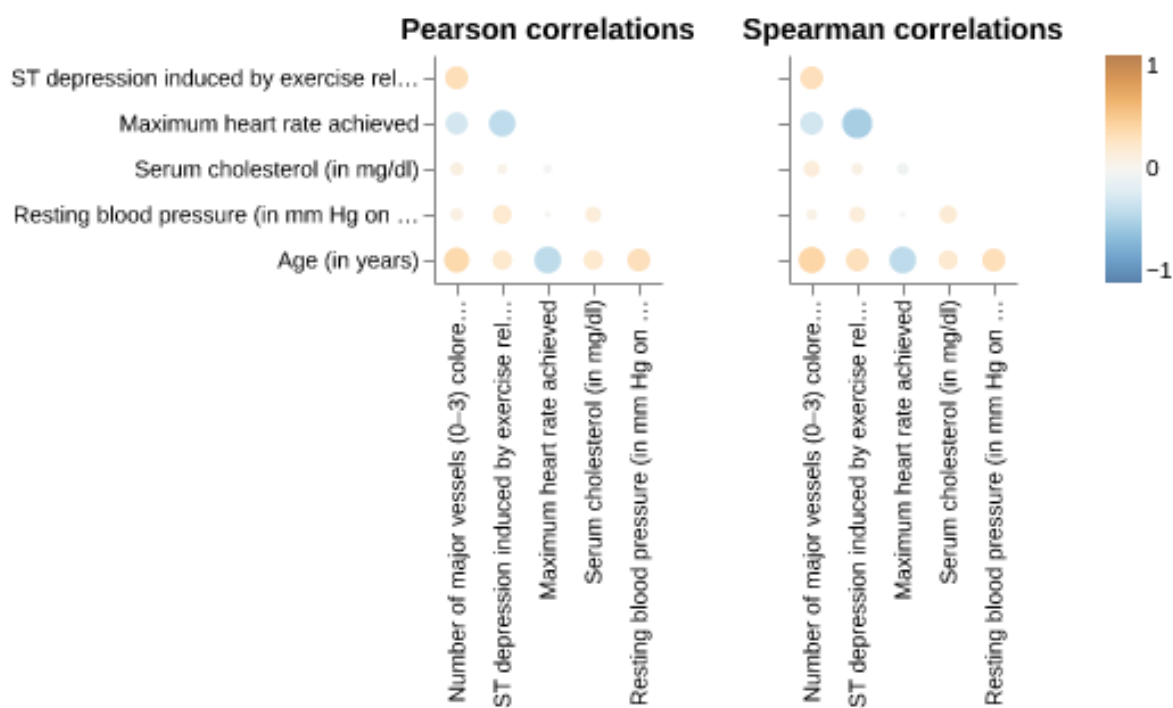
Figure 3: Correlation Matrix

For numerical feature, standard scaling is applied to keep all features within the same range.

4.2.2. MODEL CREATION

Basic models (default hyperparameter values) are now generated. A dummy model is first created to use as a baseline to use for comparison. Then, a logistic regressor and support vector classifier model is created. 5-fold cross validation (CV) is performed on each model and scores are returned.

As discussed above, F1 score is to be the primary metric to evaluate model performance, with accuracy and recall as secondary metrics. This is reflected in the scoring metrics used in CV.

Finally, a Confusion matrix was generated for each model to give an idea of false positive vs false negative rate.

4.2.3. BALANCED MODEL TESTING

Step 4.2.2 is repeated, this time creating balanced logistic regressor and SVC models, with all other hyperparameters held the same. Confusion matrices were once again generated to evaluate the types of errors seen. The goal of this step is to gain an understanding of how much accuracy is sacrificed at the benefit of improving F1 score.

4.2.4. MODEL SELECTION AND EVALUATION

With baseline models created, they are evaluated according to the criteria set above. CV scores of each model are presented below. First, standard deviation was evaluated to verify there are no abnormally performing models. Scores are presented in Table 2 below.

Comparing the metrics across models, Balanced logistic regression yields the highest recall and f1_score. For these reasons, we choose to proceed with `LogisticRegression(class_weight="balanced")`.

Table 2: Comparison of cross-validation scores across model options.

index	dummy	logreg	svc	logreg_bal	svc_bal
fit_time	0.001	0.012	0.009	0.011	0.01
score_time	0.013	0.016	0.016	0.016	0.016
test_accuracy	0.586	0.819	0.836	0.828	0.836
train_accuracy	0.586	0.862	0.915	0.866	0.919
test_precision	0	0.808	0.854	0.8	0.826
train_precision	0	0.856	0.942	0.833	0.91
test_recall	0	0.749	0.75	0.791	0.781
train_recall	0	0.802	0.846	0.846	0.893
test_f1	0	0.773	0.79	0.791	0.794

Table 2: Comparison of cross-validation scores across model options.

index	dummy	logreg	svc	logreg_bal	svc_bal
train_f1	0	0.828	0.891	0.84	0.902

4.2.5. HYPERPARAMETER OPTIMIZATION

With our final model selected, hyperparameter tuning was done via a random search methodology to find the best value for our C hyperparameter, optimizing for F1 score. The cross-validation scores among the top three models are approximately equivalent. This suggests that the model's performance is relatively stable across the parameter space, indicating that further tuning may not yield substantial improvements. With this C value selected, we proceed to evaluate the final model.

4.2.6. FINAL MODEL SCORING AND EVALUATION

With hyperparameters selected the best model is fitted on the training set, then scored on both data sets. F1 score, Recall score and Accuracy are all computed across both data sets.

Table 3: Best model metrics.

Metric	Train	Test
F1 Score	0.842	0.808
Recall	0.833	0.7
Accuracy	0.871	0.828

5. RESULTS & DISCUSSION

The model created shows great promise and with a few additional checks and improvements could be ready for deployment.

Our final test results yielded a F1 score of 0.808 and there is little discrepancy to the training F1 score (0.842). This indicates that the model does not have significant optimization bias.

Furthermore, our model has an accuracy of 0.828. This value is better than the baseline dummy accuracy, but should be higher to avoid false negatives (which can be very damaging) or false positives (which can be costly to physicians).

With hyper parameter tuning the model achieved a higher F1 score compared to original model (0.842 on train F1 score compared to 0.791 F1 score for base model cv score).
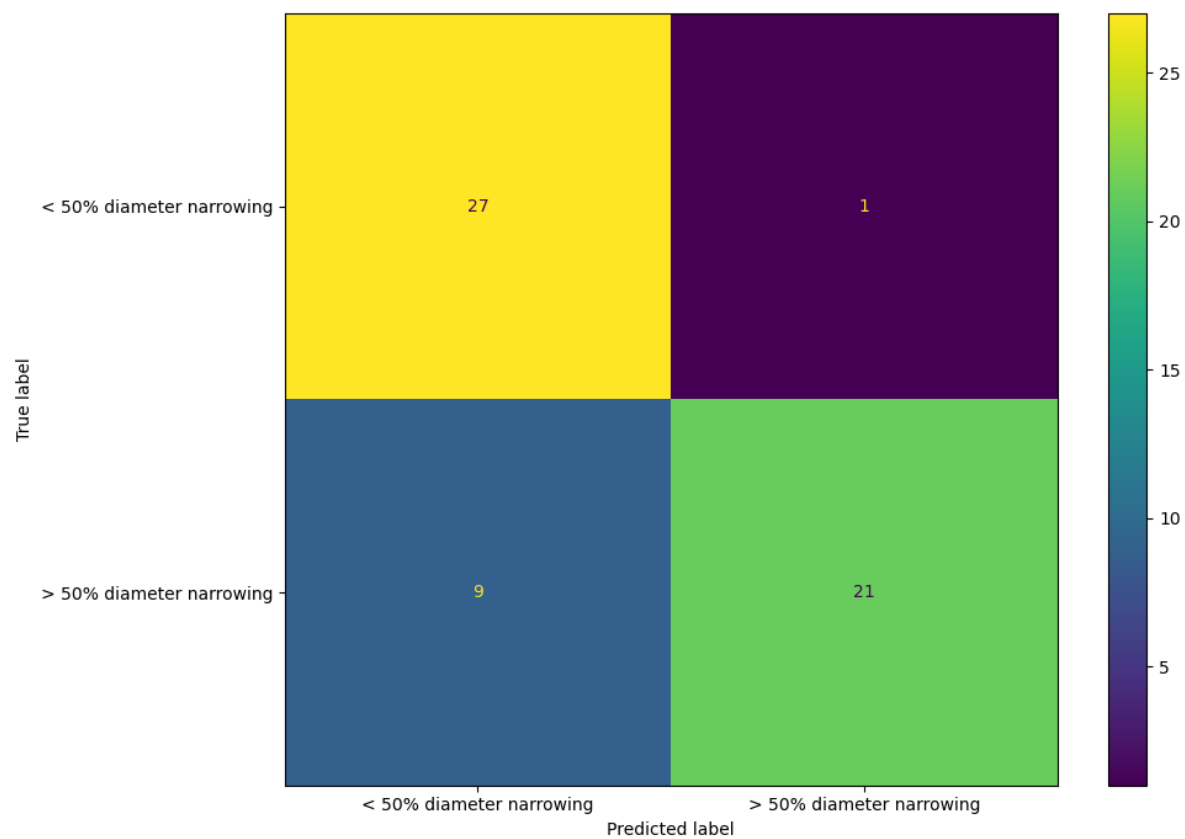
Figure 4: Confusion Matrix on Test Data

To get more rigorous performance testing and confidence in our result in future iterations, there are several improvements we can make. It is recommended to seek further data, test more model types, and conduct feature engineering.

6. CONCLUSION

The model created showed a lot of promise, being able to correctly classify presence of angiographic coronary disease with an accuracy of 0.828.

The model performed fairly well on F1 score and was able to somewhat minimize the number of false negatives classified (recall of 0.7).

There are some limitations to this report that should be noted both at the analysis level and application level.

On the analysis side, only 2 models were tested. While their performance was encouraging, a more rigorous approach would test a variety of classifiers before proceeding with logistic regression.

As well, further hyperparameter optimization could be conducted. While a wide range of C-values were tested, Only 50 possible values were tested from this range. An improvement to this would be to randomly sample from a log-uniform distribution to obtain our best C value.

On the application side, we should note that this model was tested specifically for one type of heart disease. The scope of the data used in training should be taken into account before proceeding with prediction on new data. As well this model requires a significant amount of medical information about a patient in order to create a prediction. Most of the information used to create the features to train the model is obtained through angiography, a process which itself ends in a diagnosis of the disease. So, it is worth noting that even a high performing model will not be immediately applicable, though it gives confidence on the process.

Overall, we recommend further pursuing optimization of this model. While initial results are promising, we would strongly recommend performing further testing on the model on new, larger data sets before proceeding with it.

7. REFERENCES

Athanasios Aessopos, Dimitrios Farmakis, Maria Kati. 2007. "Heart Disease in Thalassemia Intermedia: A Review of the Underlying Pathophysiology." https://haematologica.org/article/view/4438.

Janosi, Steinbrunn, Andras, and Robert Detrano. 1989. "Heart Disease." UCI Machine Learning Repository.

Master, Arthur M. 1969. "The Extent of Completely Asymptomatic Coronary Artery Disease." https://doi.org/10.1016/0002-9149(69)90064-2.

O'Flaherty, M, E Ford, Steven Allender, P Scarborough, and S Capewell. 2008. "Coronary heart disease trends in England and Wales from 1984 to 2004 : concealed levelling of mortality rates among young adults," January. https://dro.deakin.edu.au/articles/journal_contribution/Coronary_heart_disease_trends_in_England_and_Wales_from_1984_to_2004_concealed_levelling_of_mortality_rates_among_young_adults/21047827.

Organization, World Health. 2024. "The Top 10 Causes of Death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.