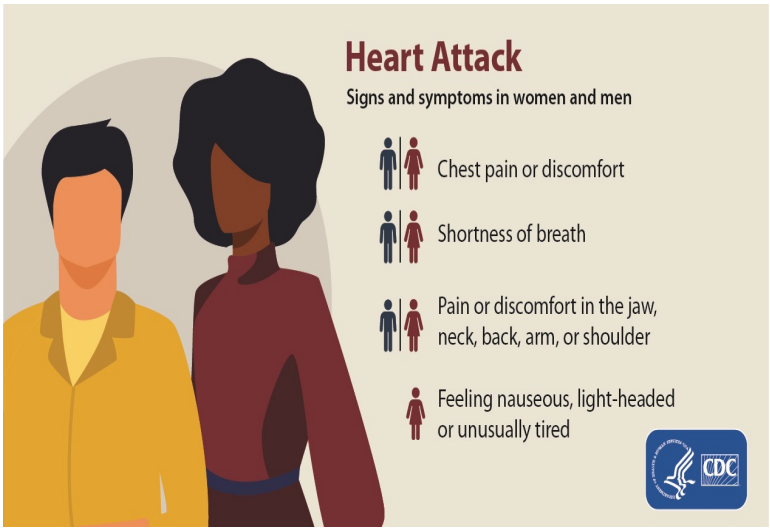
Yassine Raouz
SpringBoard
October 2022

Heart Attack Prediction

Final Report

Introduction

Heart attacks are one of the main causes of death in the world. In 2019, approximately 18 million people died from Cardiovascular diseases, representing 32% of global deaths that year. Heart attacks and strokes represent over **85%** of those deaths. But what is a heart attack? During its occurrence, a lack of blood flow causes the tissue in the heart muscle to die. It occurs when the flow of blood to the heart is severely reduced or blocked. In other words, it is sudden, unexpected, and usually fatal. The goal of this project is to predict if a person is at risk of having a heart attack.



In our project, we obtained data containing the records of 303 patients. This data contained the medical records of each patient in regard to certain medical test results. However, the most important piece of information we received was the output column, which refers to the 'final conclusion' about the patient, in other words, whether the patient is at risk of a heart attack or not. Our goal is to predict this outcome given future new data, which will help us address these issues sooner for the patients at risk. We all heard the expression "Prevention is better than cure", and today we are using machine learning to help us accomplish some progress in saving people's lives.

Data Wrangling

During this phase of the project, we discovered our data a little more and tried to capture any flaws in it. Fortunately, our data is complete and has no missing values in any of its features for all the patients, however, it was still as perfect as one could assume. Initially, we assumed that all our dataset was numeric and contained continuous values. After carefully reading the documentation and understanding

Dataset statistics

Number of variables	14
Number of observations	303
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	1
Duplicate rows (%)	0.3%
Total size in memory	33.3 KiB
Average record size in memory	112.4 B

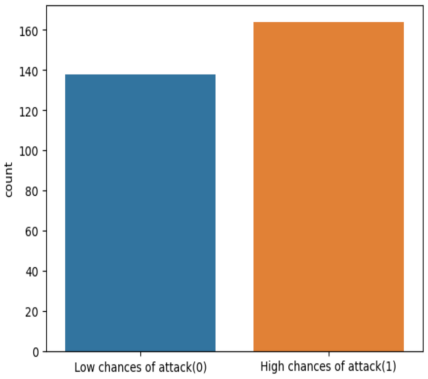
the different features, we concluded that some of them are indeed categorical rather than numeric (continuous). On top of that, we had one duplicate row, or in other words, two patients with the same results. This occurrence is most likely an entry mistake or a bug because the chances of that happening are very low.

To deal with these small issues we first looked at the unique values for the categorical features, which gave us an idea of what values are present in our data. Comparing these unique values and the documentation values allowed us to locate the “Not available” or “missing” values, which were referred to by values not present in our data description. To replace these values, and since they are categorical, we decided to replace them with the mode for that specific feature. These two features were the number of major vessels and the Thallium Stress Test result. Finally, we dropped the duplicate row and we were left with 302 patients and ready to dig deeper into our data.

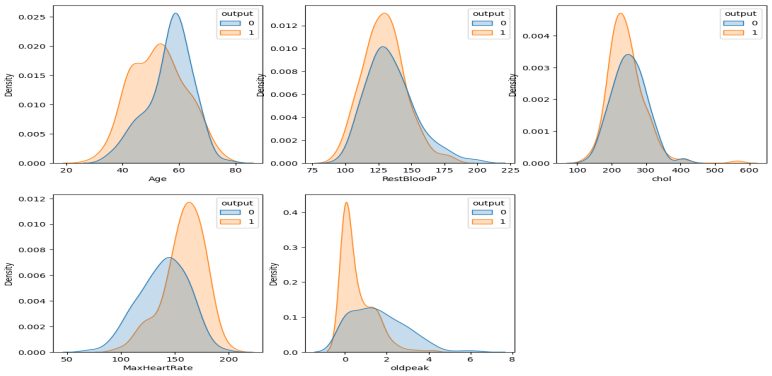
EDA

The exploratory data analysis phase was the most interesting one. Observing our features’ distribution and variance gave us knowledge about how the risk of a heart attack varies for each feature. We gained a lot of useful information to help not only help us describe our data but also to guide us in our modeling and pre-processing phases.

We tackled this phase by separating our features depending on their kind. We did not only look at each feature on its own, but we also looked at how it differs for each outcome of the target variable. Before we talk about the observations we made, we need to look at how our target variable is distributed. This count plot depicts exactly how balanced our data is, which is really helpful for the modeling phase and saves us some extra work.



1-Numerical features: A full detailed observation is present in the notebook. However, our main goal is to see how these features change our target variable. Therefore we use KDE plots and calculated the correlation between these features and the output and this is what we observed:



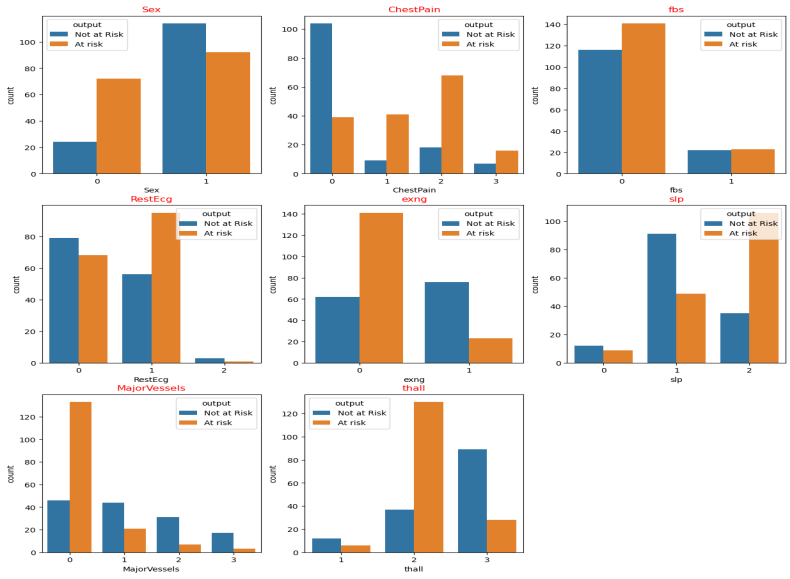
- We tend to believe as we get older we have a higher chance of having a heart attack. However, this graph shows the contrary. We notice how after 55 there is an increase in the blue graph, which refers to an increase in patients with low risk. The correlation with this variable is negative and little.

- The higher the max heart reach, the higher the chances of having a heart attack. There is some correlation there based on the graph and also the value 0.4199, which shows a moderate correlation between the target and this feature.
- For the old peak, we can see that the patients with values between 0 and 1.5 have a higher chance of a heart attack. On top of that, the graph is somewhat separated, which indicates a correlation. The correlation value of -0.429 indicates a negative and moderate correlation.
- For cholesterol and resting blood pressure, it is clear that the graphs are overlapping throughout the distribution. Also, their correlation coefficient is low, therefore for now we assume they have such a little correlation with our target value

	output
Age	-0.221476
RestBloodP	-0.146269
chol	-0.081437
MaxHeartRate	0.419955
oldpeak	-0.429146
output	1.000000

Finally, we located outliers using boxplots. However, since our dataset is so small, we believed that leaving them is the right choice!

Categorical Features:



Our observations:

- Sex:

Women (value 0) are at higher risk for a heart attack than men. Our female patients at risk are more than double the amount at low risk. While men are the opposite, there are more men at low risk than high risk. With a value of -0.283 we can say that this feature is in a negative low correlation with the target.

- Chest Pain:

Having a value of 0 was not only the most common among our patients, but it is the one where the patients at low risk almost triple the ones at risk. While patients with other values (1,2,3) are more likely to be at risk of a heart attack as shown on the graph; the Orange bars are overpowering the blue ones. With a value of 0.432080 we can say that this feature is in a positive moderate to strong correlation with the target.

	output
Sex	-0.283609
ChestPain	0.432080
fbs	-0.026826
RestEcg	0.134874
exng	-0.435601
slp	0.343940
MajorVessels	-0.463886
thall	-0.362313
output	1.000000

- FBS:

The majority of our patients had a value of 0, in other words, they had a fasting blood sugar of less than 120 ml/dl. The risk of a heart attack is

slightly higher for these people, while the patients with value 1 are still at risk, but their data is so small that we can't make a clear assumption about them. This also shows that the output doesn't make a difference in the two categories. With a value of -0.026826 we can say that this feature is in a negative very low correlation with the target.

- RestEcg:

Resting electrocardiographic results show that patients with a value of 1 are extremely at risk, as their amount doubles the patients, not at risk. For values 0 and 2, those patients are more likely to be at low risk. With a value of 0.134874 we can say that this feature is in a positive low to moderate correlation with the target

- Exng:

Exercise-induced angina with a value 0, which means absence of pain, are more likely to be at risk of a heart attack. This is surprising because it means that exercise-related heart pain has nothing to do with the chances of having a heart attack. The bars for value 1, who are people with pain, have almost 4 times more chances of not having a heart attack than being at risk. With a value of -0.435601, we can say that this feature is in a negative moderate to high correlation with the target.

- Slope:

Patients with a slope value of 2 are in general at risk of having a heart attack. As the graph shows, the number of patients at risk is almost 3 times the number of patients not at risk. As for the other values 0 and 1, the patients are more likely not to be at risk, especially with a value of 1, their chances double. With a value of 0.343940 we can say that this feature is in a positive moderate to high correlation with the target.

- Major Vessels

The value 0 is what draws our attention in this graph. We can see that patients with 0 major vessels are 3 times more likely to be at risk than not to be. Patients with 1,2 or 3 major vessels are clearly at less risk. This is also the feature with the highest correlation coefficient in our categorical features. With a value of -0.463886, we can say that this feature is in a negative moderate to high correlation with the target.

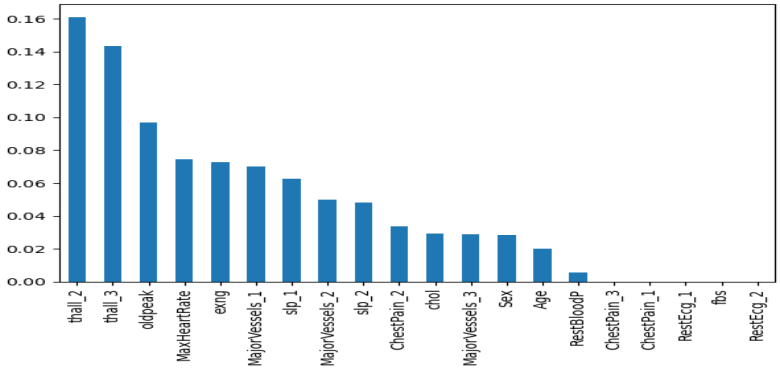
- Thall:

For Thallium stress test results, patients with value 1 have almost 4 times more chances of being at risk rather than not. While patients with 0 or 2 are more likely to be at low risk. With a value of -0.362313, we can say that this feature is in a negative moderate correlation with the target.

Pre Processing

During this phase, we made a few changes to the data to prepare it for that modeling phase. We learned about the different techniques of scaling and encoding data. For starters, we decided to encode our categorical variable even though they were numeric already and that is due to them being nominal. As for scaling, we went with Robust Scaler. The main reason for that was the presence of outliers in our dataset and we have previously made the decision not to drop them due to the small size of our data. It is important to point out that the scaling happened after splitting the data into a test and training set, and this is to avoid any data leakage.

After creating scaled and well-encoded data, we had an initial look at our feature selection technique using Mutual information (MI). Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. The results somewhat confirmed our correlation results from before as the Thallium test feature is topping the rest, while MaxHeart Rate and slope and old peak are pretty important too. The following graph shows them ranked from most important to least.



Modeling

Before modeling, we ought to talk about metrics. In our case, we will make Recall our main metric. However, we will look at other metrics to get more insight into the strength of each model and how they differ in their predictions for our dataset. Recall, also known as **Sensitivity,** focuses on minimizing the False negatives, which in our case are important. We rather predict that a person is at risk (1) even if they are actually not, rather than the opposite. Another metric we will consider is the F1 Score; it represents the weighted average of Precision and Recall. Finally, the ROC-AUC score, whose probabilistic interpretation is that if you randomly choose a positive case and a negative case, the probability that the positive case outranks the negative case according to the classifier is given by the AUC.

We tested different models with different parameters to have an idea of which ones are performing better. We could have prior assumptions that models like Naive Bayes and SVM will not outperform the other models just from experience, however, with a small dataset surprises could happen. In fact, no model is the best model, it all depends on the data in hand and its qualities and ultimately the metrics we are hoping to optimize.

After applying cross validation for Xgboost, Random Forest, SVM, Naive Bayes and Logistic Regression, we ended up with the following results.

	Algorithm	Recall train score	F1 train score	ROC-AUC train score
7	Naive Bayes	0.969231	0.819926	0.913586
1	SVM	0.908242	0.871061	0.916933
0	Logistic Reg	0.885165	0.872435	0.922527
3	Random Forest Gini	0.816484	0.818138	0.892782
2	Random Forest Entropy	0.793956	0.806526	0.901349
6	XgBoost 50	0.786264	0.796574	0.872478
5	XgBoost 500	0.785714	0.785407	0.863536
4	XgBoost 100	0.770879	0.780522	0.871129

	Algorithm	Recall test score	F1 test score	ROC-AUC test score
2	Random Forest Entropy	0.825000	0.802222	0.869444
3	Random Forest Gini	0.816667	0.791508	0.837500
6	XgBoost 50	0.766667	0.740317	0.781944
4	XgBoost 100	0.766667	0.739127	0.793056
5	XgBoost 500	0.766667	0.724841	0.781944
0	Logistic Reg	0.758333	0.793810	0.831944
1	SVM	0.750000	0.753651	0.827778
7	Naive Bayes	0.725000	0.757937	0.815278

Even though our dataset is small, Random forest and XgBoost performed slightly better in testing than the other models. However, during training, Naive Bayes, SVM and Logistic Regression had a higher recall score than the rest of the models but their test scores were a lot lower, which indicates the presence of overfitting in their training process. Xgboost performed slightly better on training than testing, which is normal and indicates that there was not much overfitting. Nonetheless, Random forest produced an unusual result; the recall score for testing was higher than training. This could be due to the size of our dataset, since one mistake(misplacement) made in classification has a stronger effect on the overall performance of the model. As for the F1 score and ROC-AUC, their results matched the recall for the most part and did not indicate any illogical results.

Therefore, we decided to only perform Hyperparameter tuning on Random Forest and Xgboost in the hopes of maximizing their performance, hence, returning a higher recall score. It turned out that the best performing model is Random Forest with a recall of 0.86 after cross validation on the testing set.

Conclusion & Key findings

Heart attacks are one of the most common causes of death in the world. While exploring this data, a lot of interesting facts about the people at risk rose up and were quite surprising. We usually relate heart attacks with age, chest pain and sometimes even gender, however, the data showed some interesting discoveries. Thallium stress level and oldpeak proved to be the most influential criterias to decide if a person is at risk or not; along with other factors like the number of Major Vessels, the cholesterol, Resting Blood Pressure and finally age.

For this project we explored multiple models to decide on which ones explains our data best. Our expectations were met as XGBoost and Random Forest performed slightly better than the other models even though our dataset is small. As a metric, we chose Recall since it is the best fit for situations like disease detection. After tuning our chosen model, RandomForest Classifier, we were able to slightly improve the recall score. Our model returns a score of 0.85 in recall after performing a cross validation on the testing set. Our classifier also provided a final ranking of the most important features in predicting the target variable, which to certain extent matched our expectations from studying the correlation of our different features.

For the future, I would hope that more data will be available to train our models better and allow ourselves to make stronger assumptions about the target variable. The more patients we have the more training our model will receive, which will help it predict and perform well on our test set. Also, models like XGBoost are known to outperform other simpler models, however, they need to be fed a lot more data. Therefore, we shall devote more time and money into collecting more data and investigating this tragedy which was believed to be unpredictable, but today with the help of machine learning, we can say with confidence that we are heading into the right direction to minimize heart attacks.