Capstone: Improve heart prognosis for patients

By Meryl Gabrielle Tubio 100763231

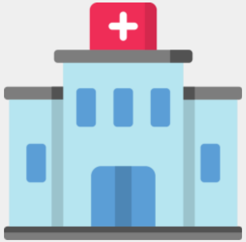
December 18, 2020

Table of Content

I. Overview

Princess Margret Hospital (in partnership with the UHN Echocardiography Lab) aims to save lives through early detection of heart murmur. With the heart murmur dataset provided by the Hospital, Researchers aim to detect early signs of heart murmur. In the hopes that this study would lead to better preparation and precaution for identified patients.

The objective of this project is to identify a pre-diagnosis of potential heart issues in potential patients. If Princess Margret can easily detect and diagnose these patients, it would result in increased awareness and prevention, and ultimately a reduced mortality rate for cardio-pulmonary patients.



"Icon made by Freepik from www.flaticon.com"

Key Questions

Key Questions and Scorecard

1. What features are important in the model?

2. What is the criteria for selecting these features?

3. How to evaluate the model performance?

Scorecard				
Perspective	KPI	Measure	Threshold	Priority
Model Performance	Overall model accuracy	Precision Recall	87%	H
Feature Importance	Permutation importance	Ranking	n/a	H

Data Analysis

Data Preparation

1 Understand the Raw Data

2 Pre-process the Data

3 General Insights

The following are known characteristics of the data:

- ✓ Categorical data
- ✓ 5011 patients
- ✓ 40 heart valve measurements (x1 to x40)
- ✓ 450 missing values
- ✓ 9 duplicate rows
- ✓ Class 0 – No Heart Ailment
- ✓ Class 1 – Heart Ailment (Congenital Defect or Heart Valve Defect)
- ✓ Represents approximately 10% of the current dataset

Data Preparation

1 Understand the Raw Data

2 Pre-process the Data

3 General Insights

Pre-process the data:

1. Identified and removed 450 missing values
2. Identified and removed 742 Outliers care of Turkey Method[1]
3. Identified class imbalance and performed class balance care of Synthetic Minority Oversampling Technique (SMOTE)[2]. From an original class ratio of 1:2 (Class 0 - 1438: Class 1 - 2820) to an improved class ratio of (Class 0 - 2256: Class 1 - 2256)

Data Preparation

1. Understand the Raw Data
2. Pre-process the Data
3. General Insights

Lastly, the following observations were evident in the dataset when the dataset was visualized:

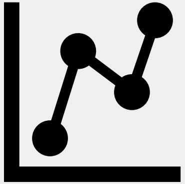
1. Extreme fluctuations in correlations from features X1 to X20, feature from high to lows correlations. While all other features, namely X21 to X40, remain relatively low in terms of correlation.

2. The dataset is normally distributed. Although there seems to be a sharp spike in 0.40 and -0.5 measurements in the dataset.

Assumptions and Constraints

1. The data has been vetoed and cleaned

2. The data is based on valid individual patient data



3. The data is normally distributed

Constraints	Type
1. The dataset cannot be increased	Hard
2. The dataset cannot have additional features	Hard
3. Limited knowledge of the dataset	Soft

Model Analysis

Logistic Regression | Decision Tree | Random Forest

Logistic Regression is a linear model that is designed for binary variables. For this report, the logistic regression is the base model used to compare the other models, primarily due to its performance in the learning curve[1] and metrics. Logistic regression has

The Logistic Regression learning curve indicates a good bias-variance trade-off, see figure 1. In addition, Logistic regression assumes the following characteristics of the data:

1. Binary – Heart murmur is binary (Class 0 and 1)
2. A linear relationship between the independent variables and the link function[2]
3. The dependent variable ("Class") is mutually exclusive and exhaustive

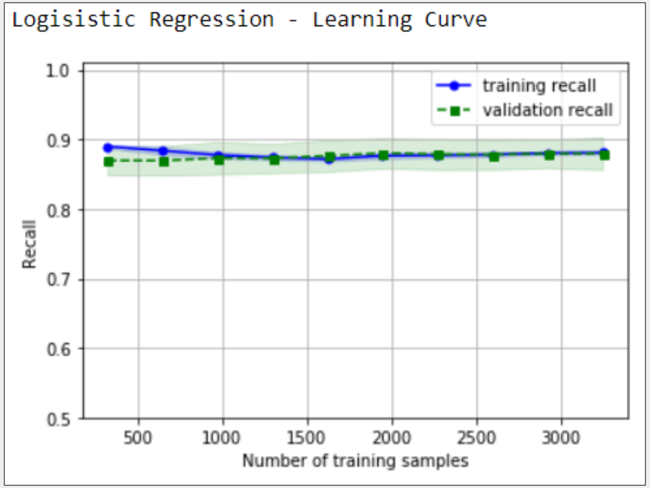


Figure 1

Model Analysis

Logistic Regression | Decision Tree | Random Forest

To focus more on important features, Researchers have decided to utilize key features identified by the SelectFromModel[1] function, Figure 2 below shows the results of this function. ***Moving forward, the results are solely based on the attributes indicated below for the 5512 patients.***

```
Key Features: Index(['x3', 'x4', 'x5', 'x6', 'x7', 'x9', 'x10', 'x11', 'x12', 'x13', 'x15',
       'x16', 'x17', 'x18'],
      dtype='object')
```

Figure 2

Model Analysis

Logistic Regression | Decision Tree | Random Forest

A decision tree is a type of supervised learning algorithm that can be used for both regression and classification problems. It works for both categorical and continuous input and output variables.

Unlike the logistic regression this algorithm does not consider any assumptions about the data and requires minimal data pre-processing. Although one drawback of this algorithm is that it is prone to overfit. The result of the decision tree is in figure 1, we can identify that there is high variance in the model.

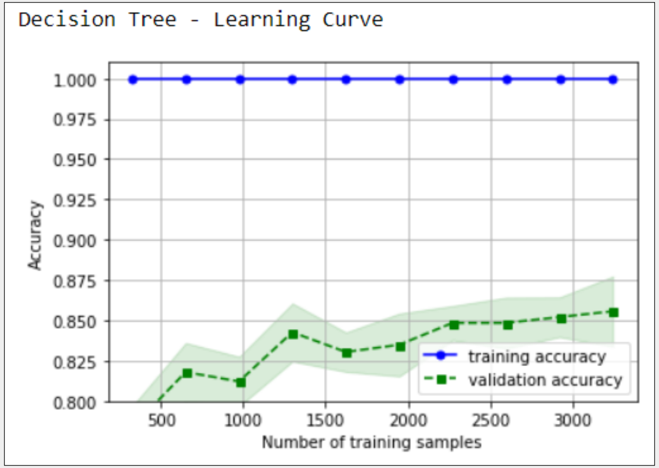


Figure 1

Model Analysis

Logistic Regression | Decision Tree | Random Forest

The random forest algorithm is a type of ensemble model consisting of multiple decisions trees. The algorithm then derives an overall output that is more accurate. Many decision trees are trained, but each tree only receives a bootstrapped[1] sample of observations. The dataset is resampled with replacement repeatedly.

Like decision tree, the algorithm has overfitted the model (see figure 1).

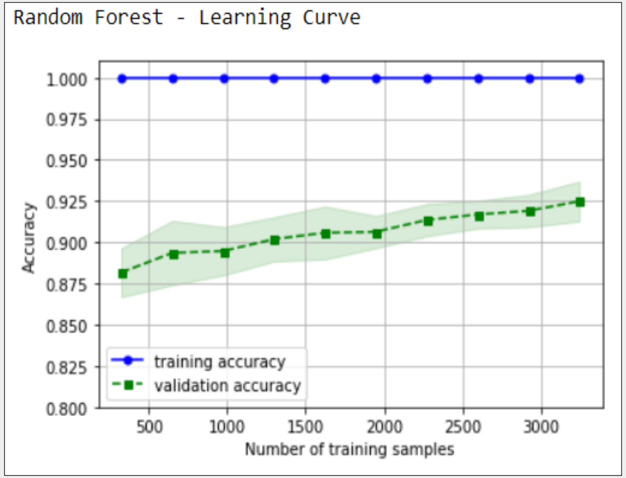


Figure 1

Model Comparison

```
[[436  16]
 [ 91 360]]

              precision    recall  f1-score   support

   Outcome 0       0.83      0.96      0.89       452
   Outcome 1       0.96      0.80      0.87       451

    accuracy                           0.88       903
   macro avg       0.89      0.88      0.88       903
weighted avg       0.89      0.88      0.88       903

NestedCV Accuracy(weighted) :0.88 +/-0.01
NestedCV Precision(weighted) :0.89 +/-0.01
NestedCV Recall(weighted) :0.88 +/-0.01
```

Logistic Regression

```
[[401  54]
 [ 75 373]]

              precision    recall  f1-score   support

         0.0       0.84      0.88      0.86       455
         1.0       0.87      0.83      0.85       448

    accuracy                           0.86       903
   macro avg       0.86      0.86      0.86       903
weighted avg       0.86      0.86      0.86       903

NestedCV Accuracy(weighted) :0.86 +/-0.03
NestedCV Precision(weighted) :0.86 +/-0.03
NestedCV Recall(weighted) :0.86 +/-0.03
```

Decision Tree

```
[[421  34]
 [ 50 398]]

              precision    recall  f1-score   support

         0.0       0.89      0.93      0.91       455
         1.0       0.92      0.89      0.90       448

    accuracy                           0.91       903
   macro avg       0.91      0.91      0.91       903
weighted avg       0.91      0.91      0.91       903

NestedCV Accuracy(weighted) :0.92 +/-0.02
NestedCV Precision(weighted) :0.92 +/-0.02
NestedCV Recall(weighted) :0.92 +/-0.02
```

Random Forest

From the 3 algorithms discussed, the random forest algorithm has the highest nested CV[1] precision and recall score at 0.92. Although discussed earlier, Random Forest is prone to overfit.

Model Comparison

Logistic Regression

Permutation Importance

Weight	Feature
0.0501 ± 0.0188	x11
0.0405 ± 0.0175	x10
0.0310 ± 0.0141	x9
0.0292 ± 0.0149	x17
0.0235 ± 0.0169	x6
0.0193 ± 0.0116	x15
0.0184 ± 0.0096	x5
0.0173 ± 0.0140	x12
0.0146 ± 0.0118	x13
0.0126 ± 0.0030	x16
0.0124 ± 0.0069	x4
0.0100 ± 0.0061	x18
0.0042 ± 0.0033	x19
0.0042 ± 0.0035	x8
0.0031 ± 0.0065	x32
0.0029 ± 0.0030	x26
0.0029 ± 0.0033	x22
0.0029 ± 0.0142	x7
0.0029 ± 0.0023	x3
0.0022 ± 0.0037	x39
… 20 more …	

Decision Tree

Permutation Importance

Weight	Feature
0.1132 ± 0.0255	x11
0.0709 ± 0.0157	x10
0.0611 ± 0.0133	x9
0.0538 ± 0.0142	x6
0.0383 ± 0.0154	x12
0.0312 ± 0.0118	x17
0.0175 ± 0.0051	x7
0.0168 ± 0.0078	x15
0.0151 ± 0.0080	x13
0.0146 ± 0.0098	x16
0.0137 ± 0.0069	x5
0.0078 ± 0.0054	x3
0.0073 ± 0.0033	x18
0.0058 ± 0.0055	x4

Random Forest

Permutation Importance

Weight	Feature
0.1010 ± 0.0191	x11
0.0558 ± 0.0143	x10
0.0456 ± 0.0129	x12
0.0332 ± 0.0074	x9
0.0244 ± 0.0051	x17
0.0217 ± 0.0057	x16
0.0166 ± 0.0056	x6
0.0166 ± 0.0070	x15
0.0166 ± 0.0142	x5
0.0157 ± 0.0068	x7
0.0117 ± 0.0089	x13
0.0095 ± 0.0095	x4
0.0091 ± 0.0088	x18
0.0033 ± 0.0046	x3

Comparative feature selection models

Based on the results of the permutation importance, heart measurement x11 and x10 have consistently ranked 1st and 2nd in among the three algorithms.

Key Findings

Random forest has the best results, but these results need to be proceeded with caution because of the current overfitted results from the learning curve. The Decision tree had the lowest results and prone to overfit. These first two models generalize the model, by not accounting training data. While Logistic regression had better precision and recall scores that wasn't prone to overfitting or underfitting.

Lastly, Features x11 and x10 have scaled the highest in feature importance for heart measurements.

Conclusion and Next Steps

Conclusion

In conclusion, an ensemble model like random forest performed the best amongst the three models. Although prone to overfitting, Analyst could investigate more in revising the hyperparameter tuning method of the algorithms and include more relevant training data. Features x11 and x10 have consistently been ranked as the most important feature for heart measurements. In other words, slight fluctuations in these heart measurements would be critical to a patient’s heart diagnosis.

Cannot definitively conclude the criteria for feature importance, however, there is evidence that once attribute x11 and x10 are reshuffled – this significantly decreases the predicting power.

Next Steps

1. What are the x11 and x10 heart measurements?
2. Can the hospital recreate this model in a larger scale? Since the current dataset represents only 10% of the entire data.
3. Is it feasible for the hospital to execute this model? Do they have enough resources and skillset to implement this model?

Appendix

Appendix

Appendix A

Appendix B

Thank you!