Final Report

Name: Kartik Sojitra

**Table of Contents**

* Introduction
* Problem Statement
* Key Questions with explanation
* Dataset Summary
* Model Analysis with Parameters
* Model analysis
* Alternative Model-1
* Alternative Model-2
* Key findings
* Results
* Conclusion
* Next steps.
* Appendix-I
* Appebdix-II

# **Introduction**

# “Heart Murmur” is one of the world’s largest disease as of now. It has been noticed that Heart Murmur are not diagnosed at the right time and due to which it reduces patient’s compliance.

In regards to this, the team had brain stormed and decided to do an analysis, finding out a predictive model which can identifying Heart murmur at different Stages of life, ranging from Birth till grown adults.

The project analyses the heart valve measurements using Echo cardiograph reports of 5009 patients at Princess Margret Hospital, in partnership with the UHN Echocardiography Lab.

The report analyses the dataset using detailed process of machine learning, which helps to identify and address the problem.

**Problem statement is**,

“Will the report be able to identify and develop a predictive model which can determine Children with Congenital heart defects and Adults with Heart valve defects using machine learning algorithm?”

The answers for these questions will help the audience to take it further developing a business model which can help to creating a portable device which can predict and identify Heart murmur.

Heart valve defects can be seen at the time of birth with new born babies and also it is seen in adults as well. With this project, the algorithm will find out a prediction model to diagnose heart defects in other patients across the age group.

Key Questions:

|  |  |  |
| --- | --- | --- |
| S.No | Questions | Need of the question |
| 1 | Will the model be able to identify and predict Heart Murmur in patients at different age group? | Yes, A Classification report with the confusion matrix, learning curve, box-plots are the metrics which will be used to evaluate the algorithm to find out accurate responding model. The model will look for output variable classified as 0- Normal patients, 1- Congenital heart defect-Babies, 2- Heart Valve Defect-Adults. |
| 2 | Is it that important to identify Heart Murmur? | Congenital heart defects have been left undiagnosed which have results existence of heart murmur in 40-45% of Children. |
|  | Heart valve defects in adults are also undiagnosed which results heart murmur in 10% of Adults. |
|  | Early diagnosis of heart defects helps to reduce 90% of heart episodes. So, by analysing the data, the project can identify a pattern using classification model, to predict heart valve defects |
| 3 | Accuracy above 90%, will that help developing a perfect predictive model? | Yes, the metrics which the report looks at are, Precision, Recall and Accuracy. As it is medical data, higher accuracy always gets you’re the better model with less errors, to be noted, predictive model for medical data are expected to have less errors. So, 90% accuracy can help developing a predictive model. |

Dataset Summary

* As mentioned, the dataset is the Heart valve measurement extracted from Echo cardiograph reports of 5009 patients at Princess Margret Hospital, in partnership with the UHN Echocardiography Lab.
* There are totally 40 Independent variables predicting one output variable, “Class” which is classified as, 0, 1 and 2.
* The dataset has been explored to check the size and shape and also how well it is distributed.
* Initially the dataset was assumed not to have any null values which in turn can reduce the accuracy. But there are null values identified and which was removed.
* Same way, there are outliers which was also identified and removed to reduce the bias.
* The dataset was split as 80% train and 20 % test set.
* The dataset was not equally balanced for all 3 output variables and hence balanced it using *Smote* technique.
* Later, the dataset was explored to check if there any inter correlation >80% and can identify that none of the 40 independent variables was intercorrelated.
* Ran box-plot, q-q plot to identify how all the variables are distributed and can see the output variables are normally distributed.
* Followed by, that, the dataset was split as x and y variables.
* Then, a pipeline was created to train for the run all optimized model learning curve;
* Then the right model was developed. Ran multiple models to select the right model which gets a better accuracy of 90% and above.

Ref: Please find Python code on Appendix -I

Model Analysis:

Completing the Exploration data Analysis, the report should identify the right model which can be fit getting the accurate results. Metrics which can be considered to finalize the right model will be,

* Learning Curve

Learning Curve is really important running any model to see, how the training and validation dataset looks like. It helps us to sensitize if the model is over fitting, good performance with the training and poor performance with the validation set or underfitting, poor performance with both the dataset. It is an eye opener which can tell us if the dataset is not perfect.

* Confusion Matrix

This helps to run the output variable and see how best the prediction and accuracy matches. This helps to get the classification report. Factors that predict this are, True positive, True Negative, False Positive and False negative.

* Classification Report

Classification report can be derived with the help of Confusion Matrix and it discusses on the outcome, Precision, Recall and Accuracy.

As discussed, the three outcomes which predicts the right model with its target are as follows,

* Precision

Precision is the metric which can the output variable that how many times, it was predicted correctly and the results were also the same. It is measured in percentage and the target for this dataset is >90%.

* Recall:

Recall, predicts an output variable, that how many times it came positive.

The target is >90%

* Accuracy:
* It is the average accuracy of all three output parameters and expected to be>90%.

Before running the models, the dataset was split into train 80% and 20% test set, Which further also undergone Grid search cross validation which helps to identify the right parameter which can help to contribute for the output parameter, Further to that, the dataset gets further optimized running Nested Cross validation for another 5 times, which means it splits the training dataset 5 times, and every time running the training set, it gets validated in the test environment. This way the dataset is further getting optimized to use the right parameter which can build the outcome. Then for all 3 models, a pipeline will be created which can be used across.

**Model Analysis:**

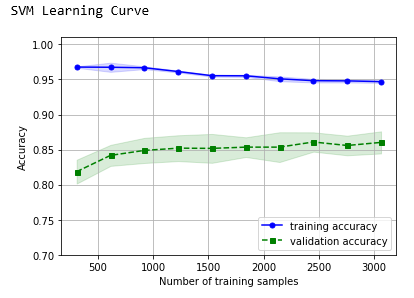
**Support Vector Machines-Base Model:**

SVM is a really good model for classification dataset. SVM operates forming a hyperplane which splits the output variable forming support vector. Distance of the nearest Support vector of each variable to the other variable’s support vector is called Margin. Higher the Margin better the model. SVM is considered to be one of the best classifier models, as it classifies each output forming the hyperplane.

The models resample the dataset and cross validates each training set with the validation set to get better accurate outcomes.

SVM results:

Learning Curve



The learning indicates that good performance with the training set accuracy, when it comes to validation set, not as good as training set accuracy. This shows that the model is overfitting with variance. However, there is an increase curve with the test which tries to be accurate. The variance has to be reduced to make the model better.

Confusion Matrix and Classification report:

* It predicted correctly 734 times out of 852 with an accuracy of 86%
* Out of 734 True positive, 237 times it predicted correctly as no Heart murmur and 244 times Congenital Heart defect and 253 -Heart Defect
* 86% of Precision and recall along 86% of accuracy

Issues:

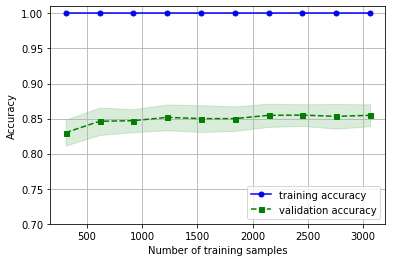
* Though SVM seems to be a good classification model, the model didn’t show good variance-bias trade off and it was over fitting. It means there is high variance, which is, the dataset is not having the precise and perfect features to predict the output which needs to addressed.
* Expected Accuracy was >90% and the model shows only 86% could be due to high variance.
* NestedCV Accuracy(weighted) :0.86 +/-0.01
* NestedCV Precision(weighted) :0.86 +/-0.01
* NestedCV Recall(weighted) :0.86 +/-0.01

Alternative model-Random forest:

Random forest is just an improvement over the top of the decision tree algorithm. The core idea behind Random Forest is to generate multiple small decision trees from random subsets of the data (hence the name “Random Forest”).  
Each of the decision tree gives a biased classifier (as it only considers a subset of the data). They each capture different trends in the data. This ensemble of trees is like a team of experts each with a little knowledge over the overall subject but thorough in their area of expertise.  
Now, in case of classification the majority vote is considered to classify a class. In analogy with experts, it is like asking the same multiple-choice question to each expert and taking the answer as the one that most no. of experts vote as correct. In case of Regression, we can use the avg. of all trees as our prediction. In addition to this, we can also weight some more decisive trees high relative to others by testing on the validation data. (Ref- <https://dimensionless.in/introduction-to-random-forest/>)

Random Forest outcomes:

Leaning Curve:



Random forest has been tried as an alternative model to SVM. Nut the results are almost the same. The validation set is not showing sign for improvement on the accuracy.

Confusion Matrix and Classification report:

* Out of 852, 736 times it shows true positive results.
* NestedCV Accuracy(weighted) :0.86 +/-0.01
* NestedCV Precision(weighted) :0.86 +/-0.01
* NestedCV Recall(weighted) :0.85 +/-0.01

Random forest was tried as an alternative to SVM, but shows almost the same results. Again, the learning curve even still worser compared to SVM.

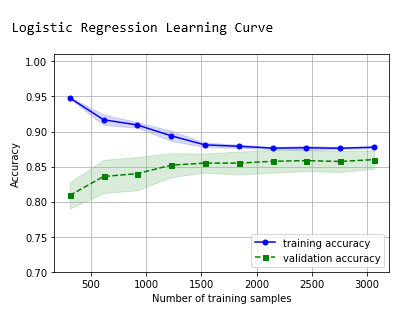
**Trying regression model as alternative:**

Logistic regression:

**Logistic regression** is a supervised **learning** classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. ... Mathematically, a **logistic regression** model predicts P(Y=1) as a function of X.(Ref- <https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm>)

Logistic Regression Outcomes:

Learning Curve



It showed a very good bias-Variance trade off as the training and validation dataset trying to travel in parallel to a common point.

This shows good performance both at training and validation set.

Confusion matrix and Classification report:

* 737 True positive out of 852.
* NestedCV Accuracy(weighted) :0.87 +/-0.01
* NestedCV Precision(weighted) :0.87 +/-0.01
* NestedCV Recall(weighted) :0.87 +/-0.01

The results 1% higher than the other model. However, by model as the nature for medical data, Logistic regression may not predict in a correct way.

Key findings:

* Though the models tried to answer the problem statement, accuracy expectation wasn’t matching.
* As discussed in model Constraints, in the last report, there is a higher variance have been noticed, which could be, the model needs more features which can contribute for the output.
* There is only the Heart valve measurement available, Patient, age, gender, race and many more details can figure out better prediction.

(More details, discussed in recommendation)

Note: Python code attached in Appendix-II

Results:

* The model predicted accuracy of 86% almost same in 3 different models.
* The issue in having less accuracy in all 3 models are same.
* But the model, didn’t stop the initiative of finding a predictive model, as it recommends for further Business case.

Conclusion:

|  |  |  |
| --- | --- | --- |
| S.No | Questions | Need of the question |
| 1 | Will the model be able to identify and predict Heart Murmur in patients at different age group? | Somehow, the model can predict different stages of Heart Murmur. However, due to lack of information, the model may not be able to tell an age range in which the disease prevails.  Discussion in the constraints, can be taken care to further more develop information’s |
| 2 | Is it that important to identify Heart Murmur? | Yes, Heart Murmur is mostly asymptomatic and cannot be diagnosed on time. 40-45 % of Children and 10% of Adults are undiagnosed with heart murmur.  Early diagnosis helps to reduce 90% of the patient life, |
|  |
|  |
| 3 | Accuracy above 90%, will that help developing a perfect predictive model? | Yes, the metrics which the report looks at are, Precision, Recall and Accuracy. As it is medical data, higher accuracy always gets you’re the better model with less errors, to be noted, predictive model for medical data are expected to have less errors. So, 90% accuracy can help developing a predictive model. |

Next Steps:

As discussed above, though SVM as a base model which is meant for Classification model, even may not produce expected results.

There may reason which results in such accuracy.

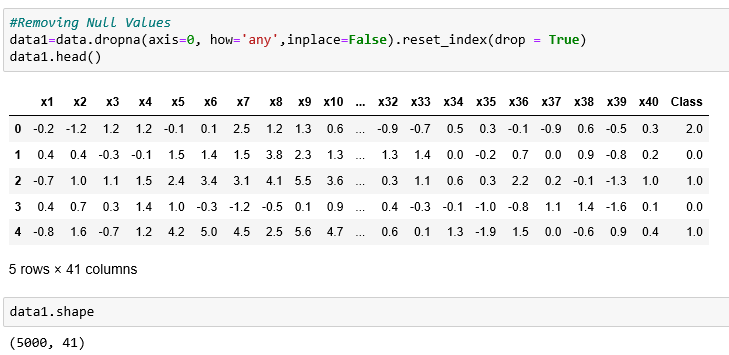
As discussed in the data set constraints. , we need few more information as below which can help to make the SVM model accurate,

* Patient Gender
* Patient race
* Patient age
* To confirm the lab reports are taken at a same centre or multi centre. If Multi-centre, need to ensure, they used the same technique for all patients measuring the valve
* Figuring out these details, we would be able to predict a model which can identify Heart murmur which has huge Business potential.

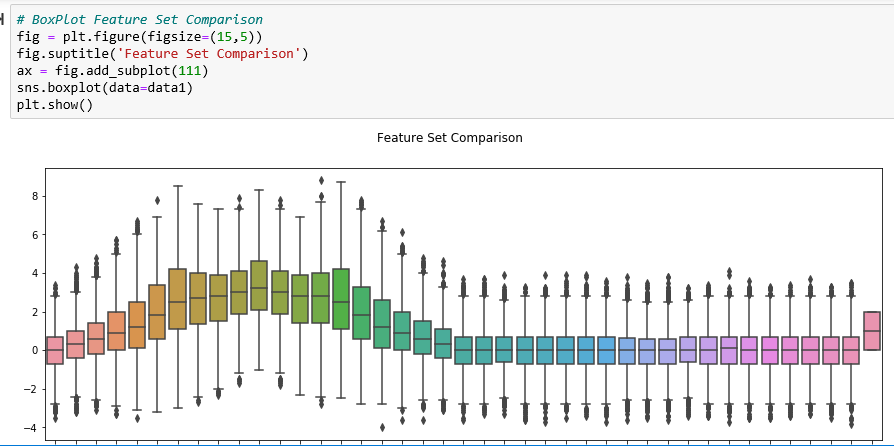
Appendix-I

Data cleaning:

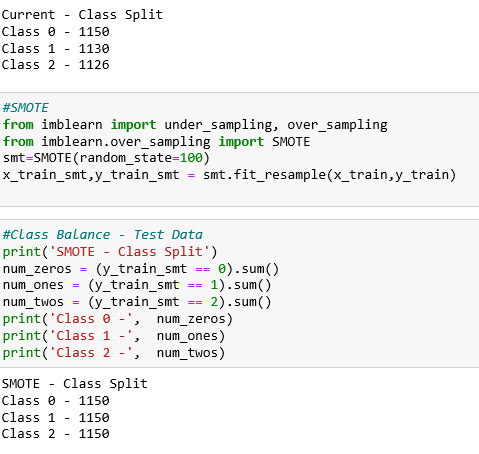
Identifying the Null values and removing it.



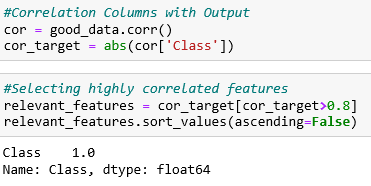
Outliers:



Balancing the dataset:

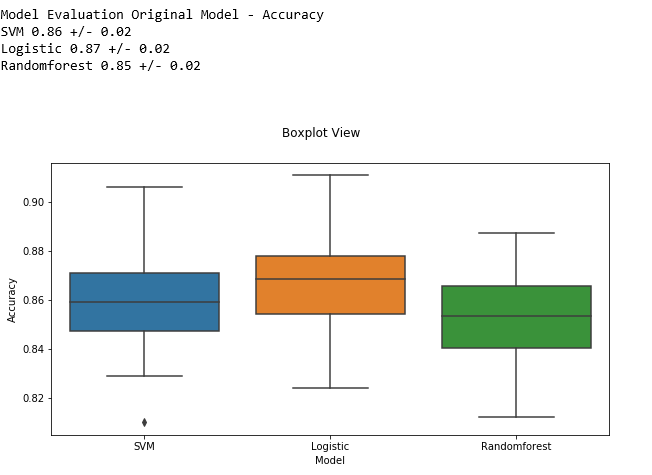


Confirming no Intercorrelation:



Appendix-II

Model accuracy after CV:



Model Outcomes:

