**AN ANALYSIS ON CORONARY ATHEROSCLEROSIS PATIENTS**

AUTHOR NAME: VIGNA SHREE TELUKUNTA

G-NUMBER: G01194577

**INTRODUCTION**

Coronary artery disease is the narrowing or blockage of coronary arteries, which is caused due to atherosclerosis. Atherosclerosis is the hardening or clogging of arteries, which happens due to buildup of cholesterol and fatty deposits on the inner walls of arteries which are called plaques. The plaques formed inside the arteries can cause abnormal tone and function in arteries or can restrict the blood flow to heart muscle by physical clogging.

When the heart doesn’t have adequate blood supply, it gets starved of oxygen and vital nutrients which are needed by heart to function properly. This could cause a chest pain which is known as angina. If the energy demand of heart becomes much greater than blood supply or blood supply to a portion of heart muscle is completely cut off, there is possibility of heart attack.

Heart disease is one of the prominent causes of death amidst men and women in the United States of America. Around 16.5 million are affected due to Coronary artery disease. About every 40 seconds, someone has a heart attack in the US, as estimated by AHA (American Heart Association). Also, for a person who has no risk factor of heart disease, have a lifetime risk of 3.6% for men and less than 1% for women, for being affected by a cardiovascular disease. If a person has 2 or more risk factors, the lifetime risk is increased to 37.5% for men and 18.3% for women, for being affected by a cardiovascular disease.

**OBJECTIVE**

To predict the mortality of patients having Coronary Atherosclerosis disease.

**DATASET**

In this project, worked on the MIMIC-III Database accessed from physionet website. MIMIC is a database containing data of patients who have stayed at Beth Israel Deaconess Medical Center in ICU (Intensive Care Unit). In the analysis of Coronary Atherosclerosis, 4 tables have been used from the MIMIC-III Database: ADMISSIONS, PATIENTS, **DIAGNOSES\_ICD and D\_ICD\_DIAGNOSES.**

**ADMISSIONS: The ADMISSIONS Table has a Unique ID (HADM\_ID), which identifies every unique hospitalization of patient in the database.**

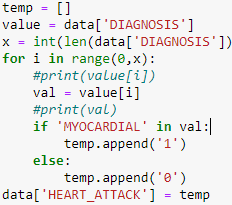
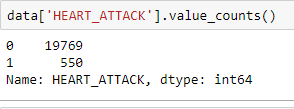
**PATIENTS: It has a Unique ID (SUBJECT\_ID) which identifies every patient uniquely in the database.**

**DIAGNOSES\_ICD: The diagnosis that are done in the hospital, are coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system.**

**D\_ICD\_DIAGNOSES: ICD Codes are assigned, which are related to procedures.**

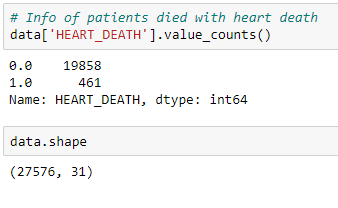
**DATA PREPROCESSING**

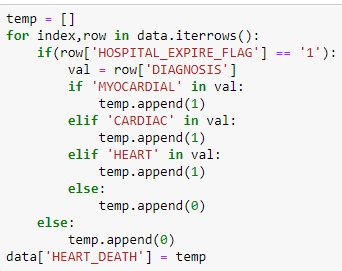
**There are many categorical variables, which have been condensed to be used for analysis. All the patients age used for analysis is less than 90.**



**Fig 2: Records of Heart Attack**

Fig 1: Collecting Heart attack data

The fig 1 shows that, by using comparison of the string ‘MYOCARDIAL’, if true we can say that the patient had a heart attack, we group all such data as HEART\_ATTACK. From fig 2, we can see that there are 550 patients have suffered heart attack.

Fig 3: Collecting Heart Death Data Fig 4: Records of Heart Death

The fig 3 shows that, by using string comparison of three key words ‘MYOCARDIAL’,’CARDIAC’ and ‘HEART’, if a record has all three words, we can say that the patient had a Heart Death.

From fig 4 we can see that 461 patients have had a heart death.

**Decision Tree Classifier**

The Decision Tree Classifier gave an accuracy of 95.7%, but the data has class imbalance problem, the precision is 0.09 and recall is 0.10. It shows an auc 52.16.

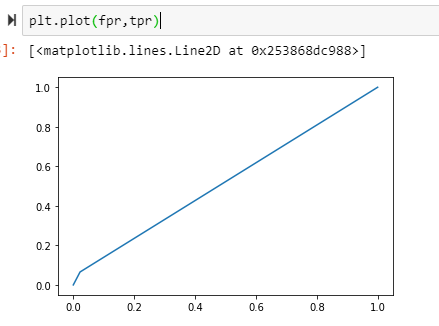
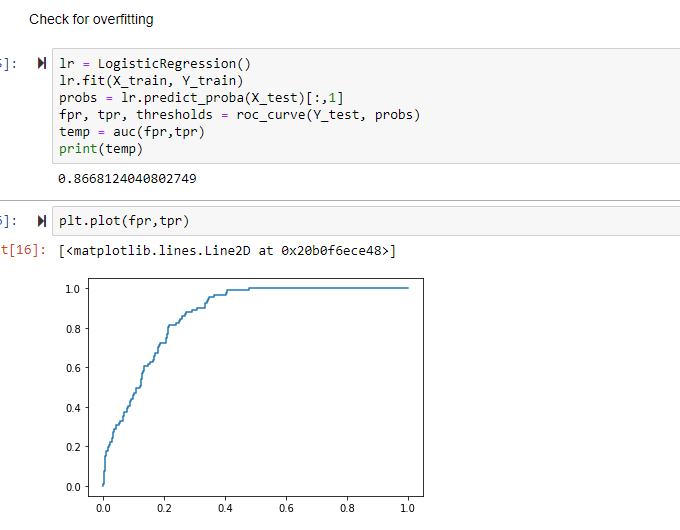


Fig 5: AUC on Test set for Decision Tree Classifier

**Logistic Regression**

The logistic regression gives an auc of 93.54 on the train set and 86.68 on the test set. This shows us that there is no overfitting on data.

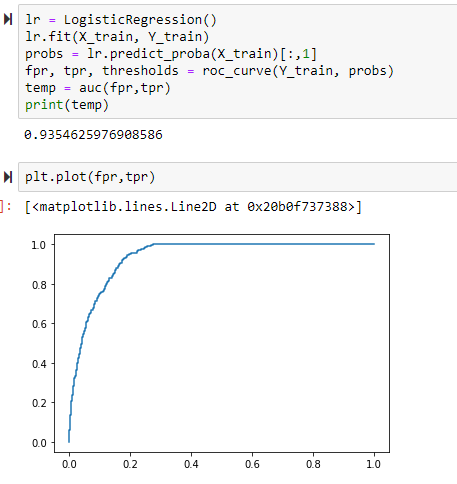
Fig 6: AUC on Test set

Fig 7: AUC on Train set

**Random Forest**

The Random Forest model gives an auc of 61.29 on the test set. The model gives an auc of 100 on the train set, which shows clearly that this model is overfitting.

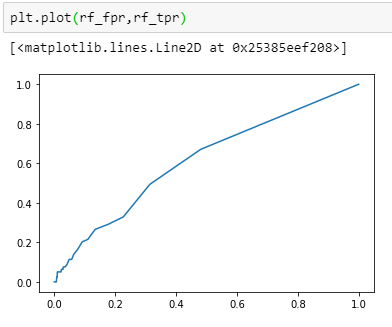
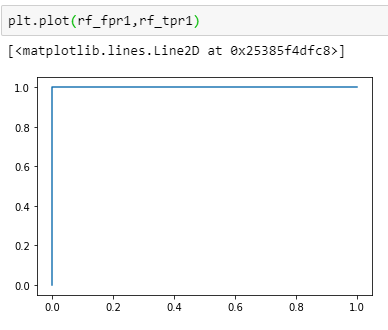
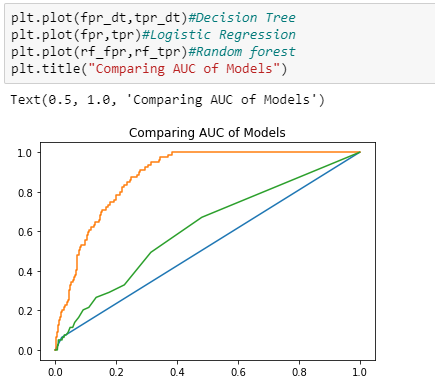
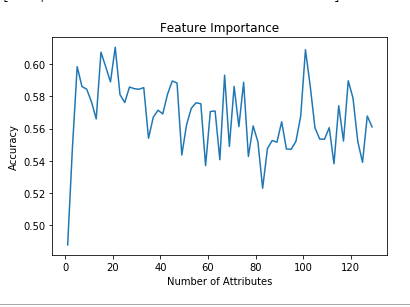


Fig 8: AUC on Test set

  
  
Fig 9: AUC on Train set

  
  
  
Fig 10: Comparing of AUC of Different Models

**Feature Importance**

Fig 11: Feature Importance Plot

The fig 11, shows the feature importance plot using Random Forest. The attributes ‘Age\_Admit’,’Age\_Death’ and ‘ADMISSIONS\_CNT’ are the important features, using these three features model can achieve upto 54.8% accuracy.

**Learning Curves**

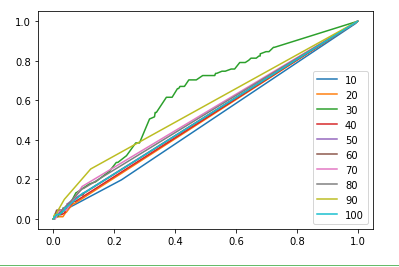


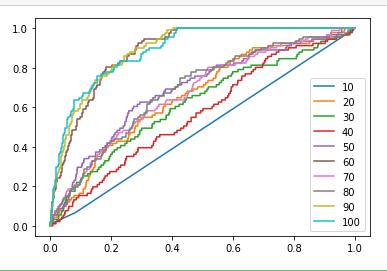
Fig 12: Learning Curve for Random Forest Classifier

Fig 13: Learning Curve for Logistic Regression

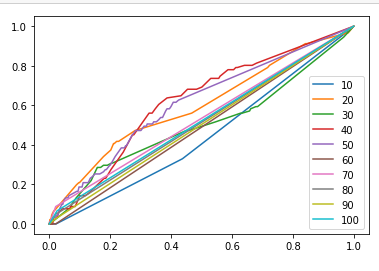


Fig 14: Learning Curve for Decision Tree Classifier

Learning curves is to plot predicted accuracy against training set size. While plotting the learning curves, we observe that as the attribute count increases, parallelly the accuracy of the plot also increases. The figures, 12, 13 and 14 show the learning curves of Random Forest, Logistic Regression and Decision Tree Classifier.

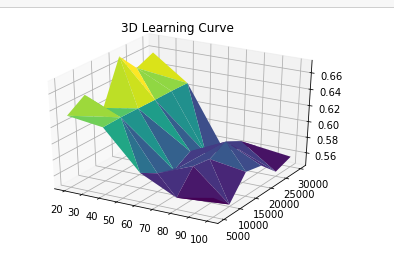


Fig 15: 3-D Learning Curve

The Fig 15, shows the 3-D Learning Curve, with number of attributes on x- axis, AUC on y-axis and size of training set on z-axis.

|  |  |
| --- | --- |
| Model | AUC |
| Decision Tree Classifier | 52.16 |
| Logistic Regression | 86.68 |
| Random Forest | 61.29 |
| Tensor Flow | 74.57 |

**Conclusion:** The best model for my data set is logistic regression, but more conclusions could be drawn from the data, if we had more records regarding coronary artery disease, as the current data is having a class imbalance problem.

From the Feature importance plot we can say that all the features are important in making predictions or we have to find the correct combination of features rather based on ranking.

**References:**

[1] Rjgiedt. “Rjgiedt/Cardiovascular\_Death\_Prediction.” *GitHub*, github.com/rjgiedt/Cardiovascular\_Death\_Prediction.

[2] Johnson, Alistair, et al. “MIMIC-III Clinical Database.” *MIMIC-III Clinical Database v1.4*, 4 Sept. 2016, physionet.org/content/mimiciii/1.4/.