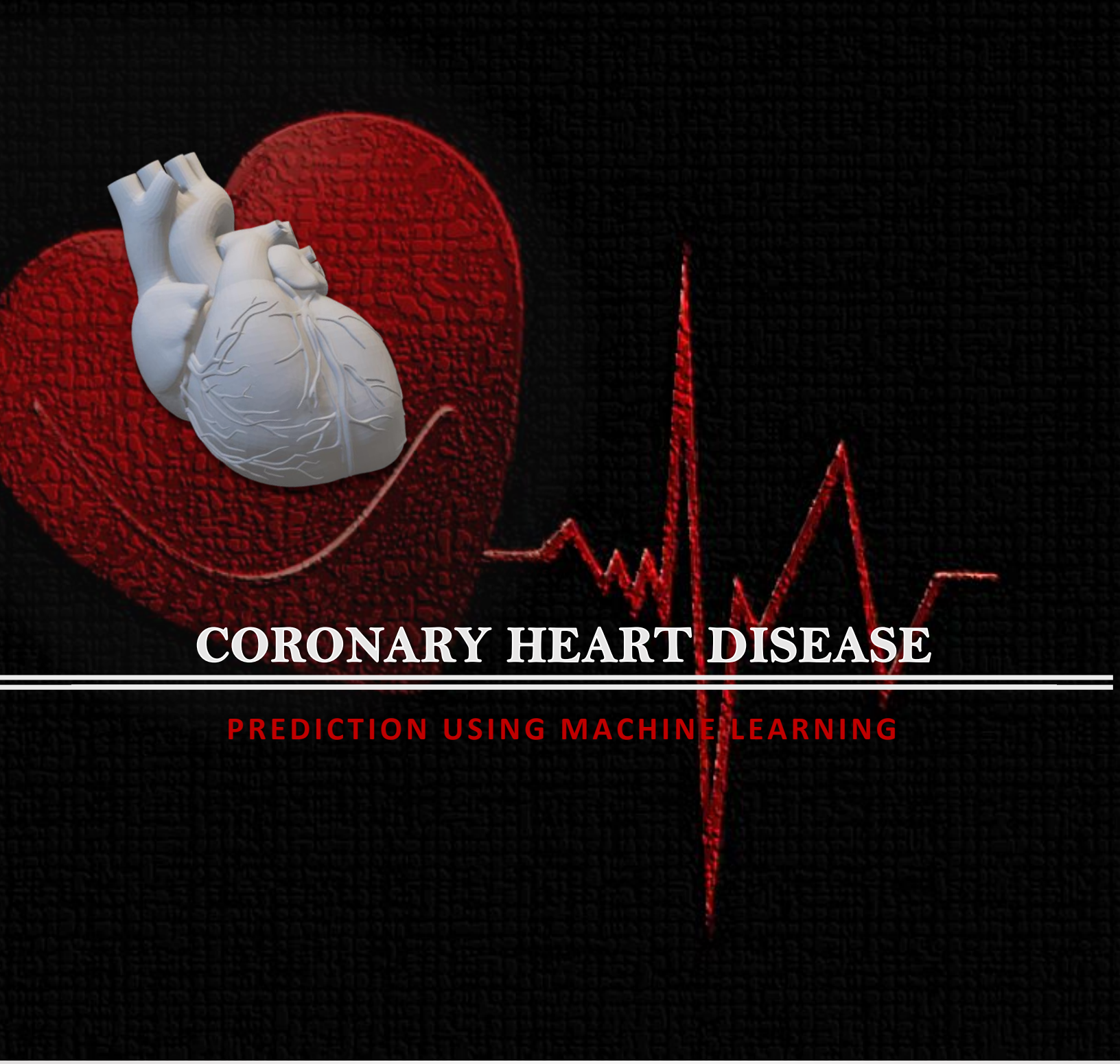


CORONARY HEART DISEASE

PREDICTION USING MACHINE LEARNING

Piya, Sarbottam (Sarbo)

CAPSTONE 2 | APRIL 14, 2021

SUMMARY

Heart disease is one of the most prominent causes of death, causing more than 17 million deaths worldwide each year.

Machine learning could be efficiently used to decode hidden information in the clinical dataset and predict the likelihood of heart disease occurrence.

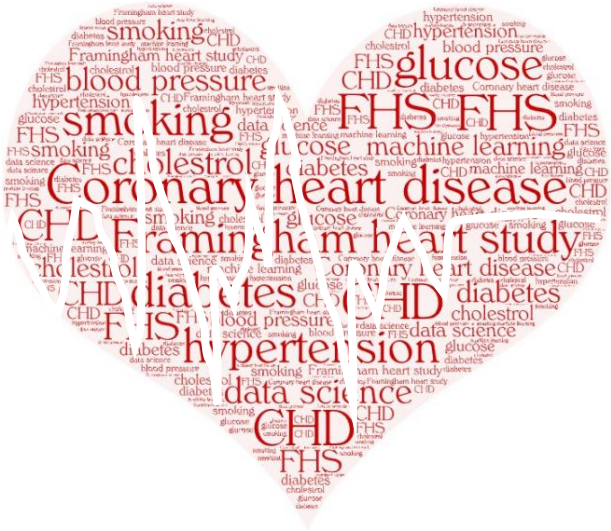
The goal of this project was to develop a model that detects possible coronary heart disease (CHD) using machine learning based on clinical data obtained from the original cohort dataset from the Framingham heart study.

Logistic regression was found to perform the best among various classifiers based on ROC-AUC score. For a patient or a healthcare professional, it is more important to have high recall value that correctly predict the likelihood of occurring CHD than correctly predicting the likelihood of not having CHD. However, the recall value of the logistic regression model was very low (0.07) which could be due to the imbalanced dataset used in this study. Probability threshold that assigns data to different classes was tuned to improve the recall (0.73; beta = 2). However, the performance of the model improved further when logistic regression model was used in oversampled data with Synthetic Minority Oversampling Technique (SMOTE). This model was able to correctly predict 88% of the positive cases.

TABLE OF CONTENT

Introduction

Heart disease is one of the most prominent causes of death around the world. According to the World Health Organization, 17.9 million people die annually of heart related diseases world-wide (World Health Organization, 2017). In the United States alone, about 655,000 people die from heart disease each year (CDC, 2020). To date, several risk factors associated with heart diseases have been identified, including: high blood pressure, diabetes, smoking/tobacco use, obesity, physical inactivity, high cholesterol level (or other lipids) and kidney disease. Since heart disease is associated with several contributory risk factors, it is often difficult to diagnose heart disease on time. In addition, the diagnosis of heart disease involves a complex combination of clinical and pathological data resulting in a very high medical cost.



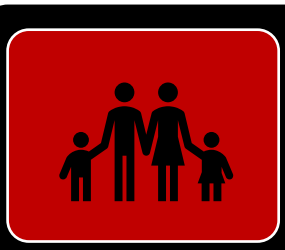
Problem Statement

Medical professionals in hospitals and various institutions, all around the world, collect data on various health related issues including heart disease. These institutions have massive medical records which are often very noisy. As a result, these huge datasets are almost impossible for the human mind to comprehend. Therefore, these datasets with a lot of hidden information are mostly ignored and clinical decisions are made based on doctors' intuition and experience. Nevertheless, these datasets can be analyzed using various machine learning techniques that allow to develop predictive models for the presence or absence of heart related diseases accurately. Therefore, the objective of this project is to develop a predictive model of heart disease using machine learning based on the demographic, comorbidity, vital sign and some laboratory investigation data. Such studies, if able to predict heart disease on time, will be useful for healthcare professionals for accurate heart disease diagnosis and treatment.

DATASET

For this project, I used the **original cohort dataset from Framingham heart study (FHS)** which is publicly available in Kaggle (https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset) in CSV format. The goal of the FHS was to identify the characteristics that are responsible for cardiovascular diseases in humans. The dataset includes 4,240 records and 16 attributes. The attributes include our target response feature-presence of heart disease (TenYearCHD).

PREDICTORS OF CHD



Demographic information

- Gender, Age And Education

Comorbidity

- Blood Pressure, Stroke, Hypertension, Diabetes, Smoking Habit

Vital Statistics

- Systolic Blood Pressure, Diastolic Blood Pressure, Body Mass Index (BMI), Heart Rate

Lab Investigation

- Total cholesterol level
- Glucose level

DATA WRANGLING

Data comprised of 16 features including the target feature which is TenYearCHD (ie. occurrence of heart disease) and 4240 rows. This was a fairly clean dataset. In data wrangling step, I performed following tasks:

1. *Checked if there were any duplicate rows but found no duplicate rows*
2. *Renamed the column name 'male' to 'gender'.*
3. *Columns 'BMI', 'cigsPerDay', 'totChol', 'BPMeds', 'education', and 'glucose' had missing data. Therefore, I deleted all the rows that had missing data. After removing columns with missing values, the dataset comprised of 3658 rows*
4. *Visualize the histogram of all the features (**Figure** 1).*

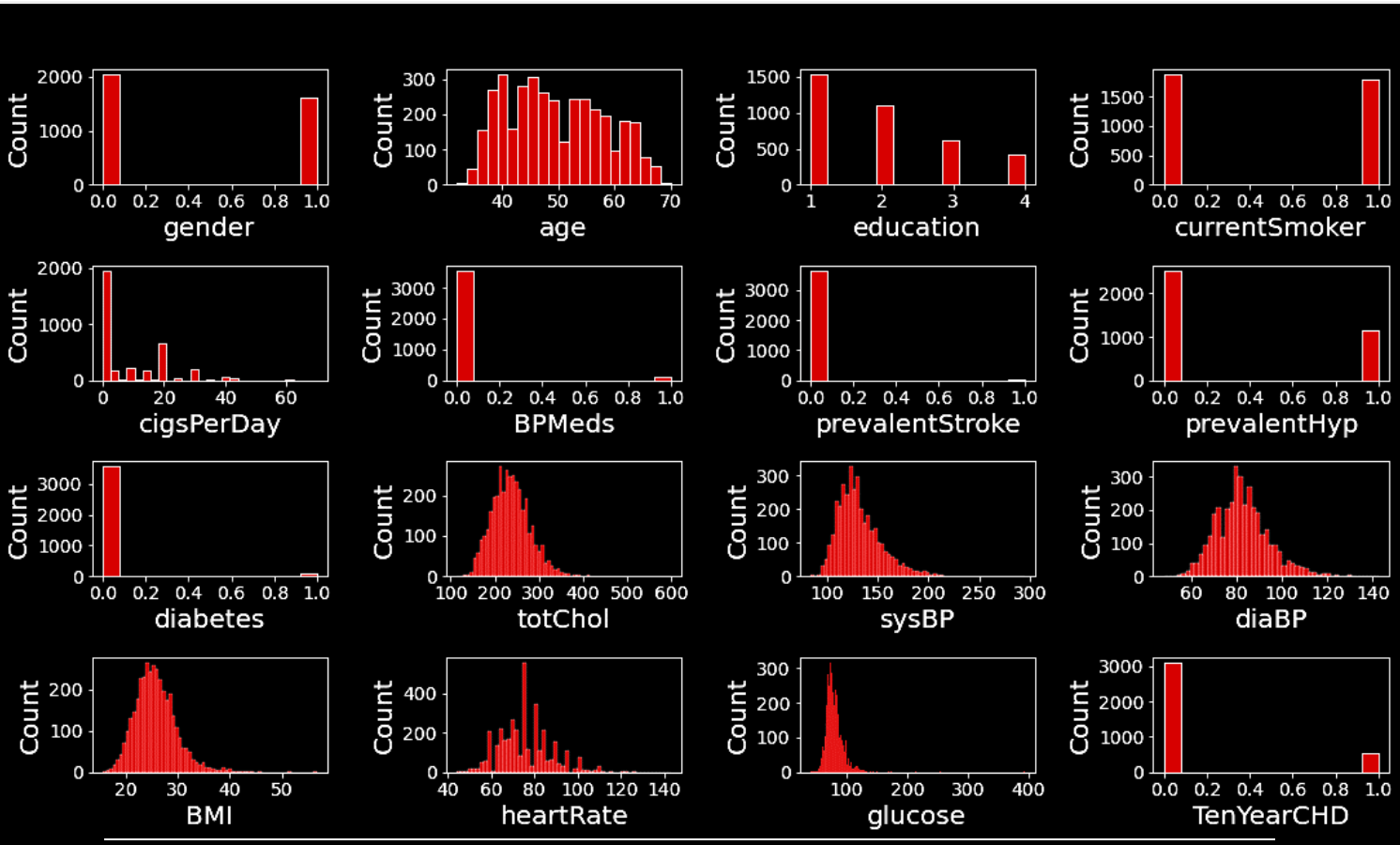


Figure 1. *Histogram displaying the frequency distribution of various features associated with CHD.*

EXPLORATORY DATA ANALYSIS

Target variable

Our target variable is TenYearCHD which represents occurrence of heart disease. We can clearly see the class imbalance in the target variable. There are 3101 cases that have no CHD while the number of cases with CHD is 557 (**Figure** 2).

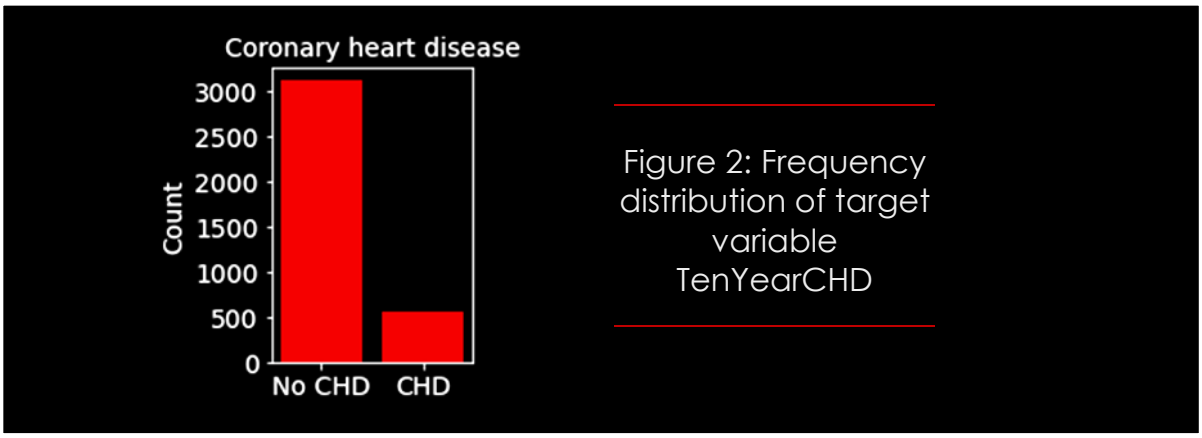


Figure 2: Frequency distribution of target variable TenYearCHD

Categorical features

The dataset consisted of six categorical features that included 'education', 'currentSmoker', 'BPMeds', 'prevalentStroke', 'prevalentHyp', and 'diabetes'. I performed a chi-square test to evaluate if there is any association between the above-mentioned categorical features and CHD and found that all the categorical features except 'currentSmoker' were significantly associated with CHD. The occurrence of CHD was higher in males than in females, higher in people with some high school education than people with higher school or higher-level education, and higher among people who take blood pressure medicine, have had a stroke, hypertension or diabetes (**Figure** 3).

Numerical features

The dataset consisted of eight numerical features that included 'age', 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'glucose'. I performed t-tests to assess if there were statistically significant differences in the means of various numerical features between people with or without CHD. The analysis showed that the average age of people with CHD was higher than that of people without CHD.

Similarly, people with CHD smoked significantly more cigarettes per day, had significantly higher average cholesterol, systolic blood pressure, diastolic blood pressure, BMI, and glucose level than people without CHD (**Figure** 4).

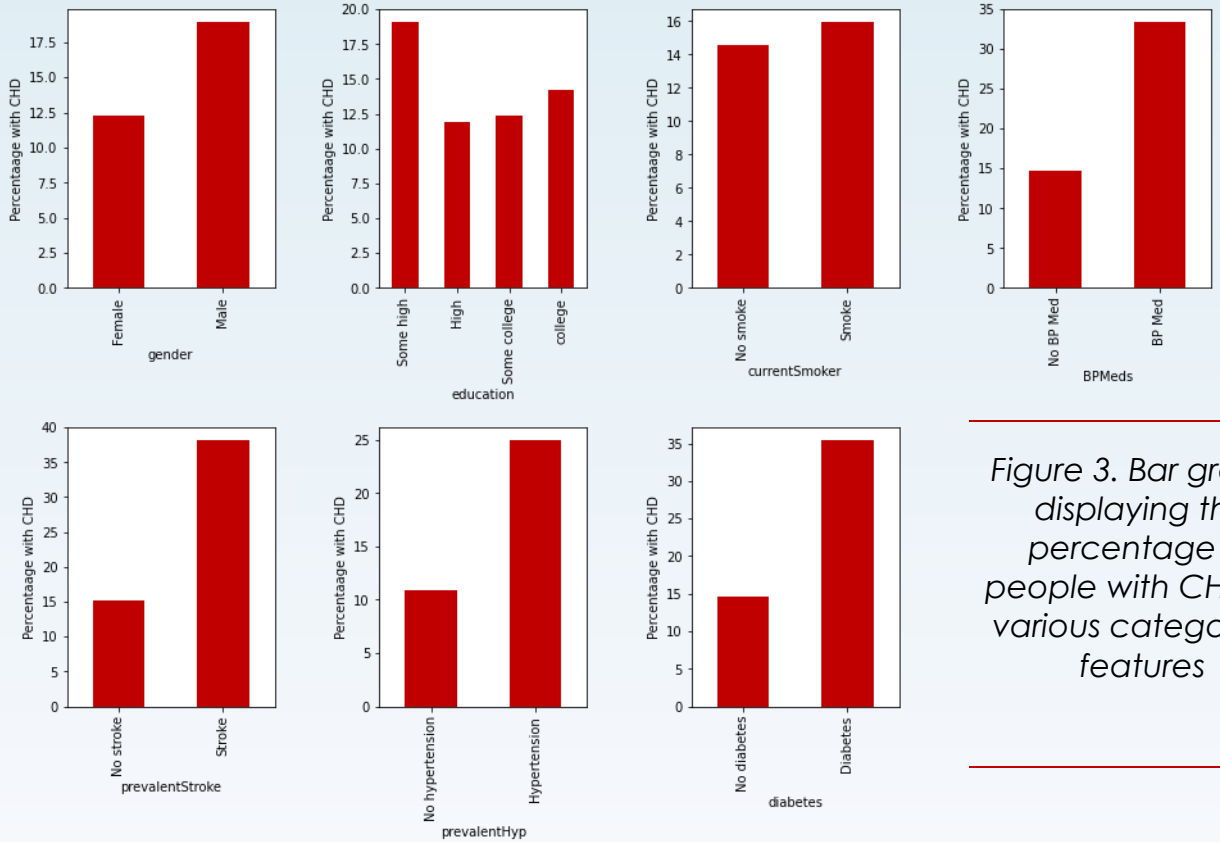


Figure 3. Bar graphs displaying the percentage of people with CHD for various categorical features

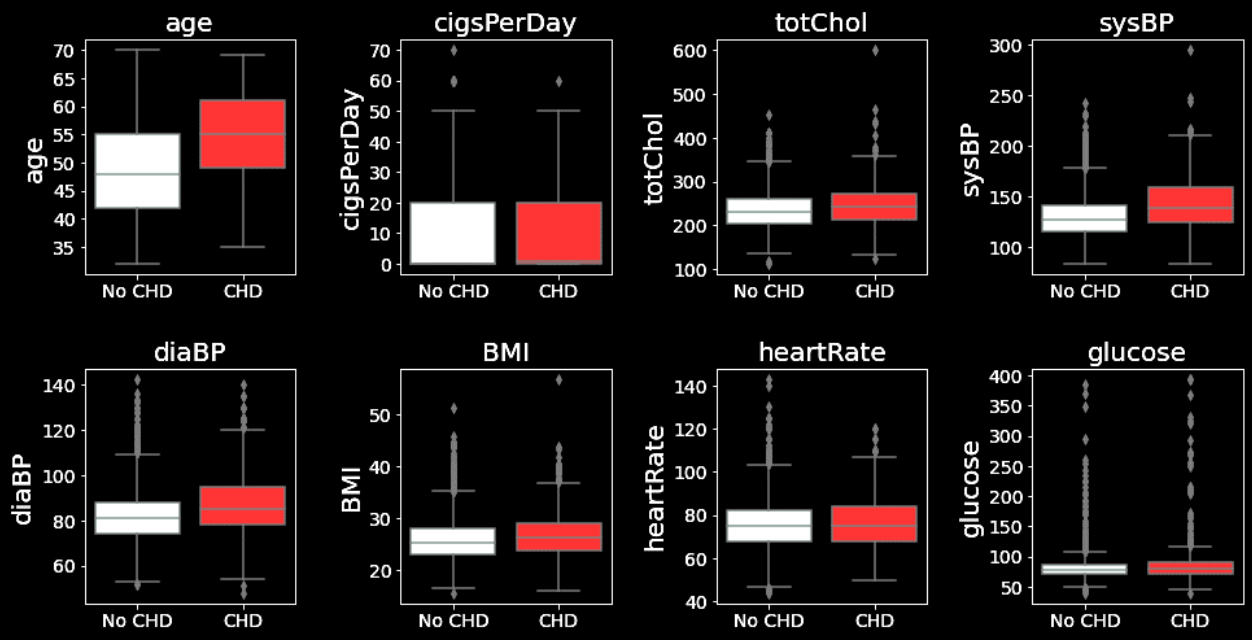


Figure 4. *Box plot displaying the distribution of various numerical features for people with or without CHD.*

Correlation among variables

The data do not show a strong correlation of TenYearCHD with any other variables (Figure 5). This may indicate that heart disease is the result of several contributory factors. There are some variables that show moderate to strong correlations. For example, diaBP, sysBP and hypertension show strong correlation among themselves. Similarly, diabetes and glucose also share a moderate correlation.

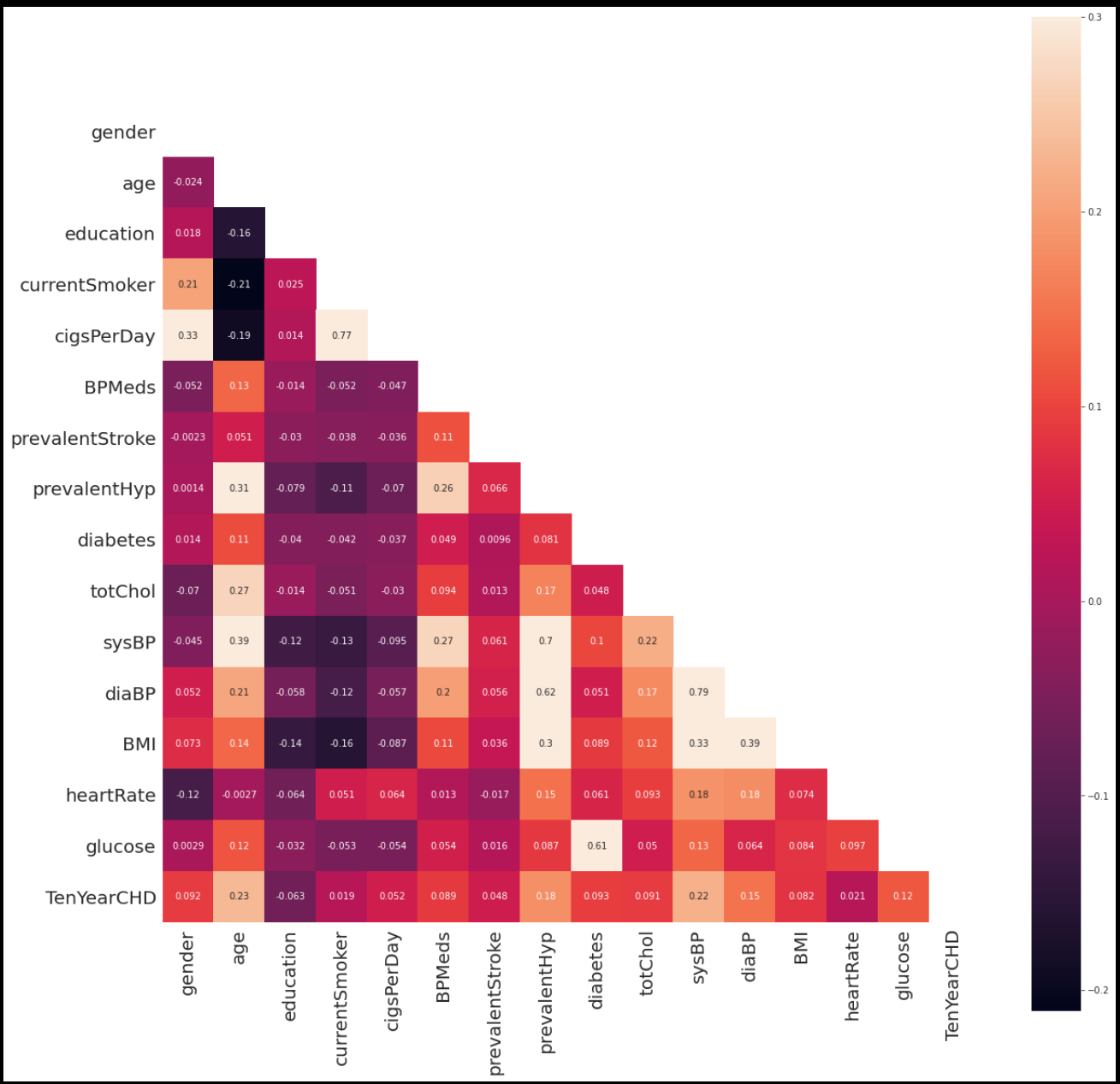


Figure 5: *Heatmap demonstrating correlation among different features*

FEATURE IMPORTANCE

Feature importance using logistic regression

Before venturing into modeling, it's important to understand feature importance. This step provides a more precise understanding of which features most affect CHD (while controlling for the others, and can be used to inform feature selection which has the potential to reduce both overfitting and training time.

When the independent variables show a strong correlation among themselves, there is probability of occurrence of multicollinearity. In order to check multicollinearity, I computed Variance Inflation Factor (VIF). As a rule of thumb, VIF of more than 5 indicates multicollinearity. However, there were no features with VIF greater than 5 suggesting no occurrence of multicollinearity (Table 1).

Provided that the input variables are in the same scale or are scaled before fitting a model, the regression coefficients can be considered an indicator of feature importance.

Table 1: Variance Inflation Factor (VIF) score of the features

Feature	VIF
Constant	3.496098
Gender	1.203755
Education	1.053926
Currentsmoker	2.583798
Cigsperday	2.730844
Bpmeds	1.111055
Prevalentstroke	1.017441
Prevalenthyp	2.048857
Diabetes	1.615771
Totchol	1.111191
Sysbp	3.730424
Diabp	2.989243
Bmi	1.239636
Heartrate	1.093428
Glucose	1.637372
Age_group	1.317082

Logistic regression analysis showed that six features (gender, age_group, sysBP, cigsPerDay, totChol and glucose) contribute significantly to CHD ($p<0.05$). Then, I computed the odds ratio by exponentiating the coefficient (**Figure** 6).

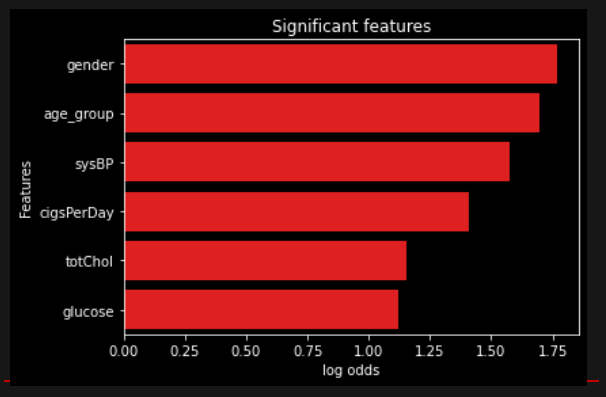


Figure 6: *Odd ratio of features that contribute significantly to CHD*

Feature importance using random forest

In the next step, I performed feature selection using sklearn random forest feature_importances_ function that helps to identify the features that contribute the most to predicting the target variable.Feature selection identified sysBP, BMI and diaBP as the top three most important features. **Figure** 7 demonstrates the rank and the scores of various features.

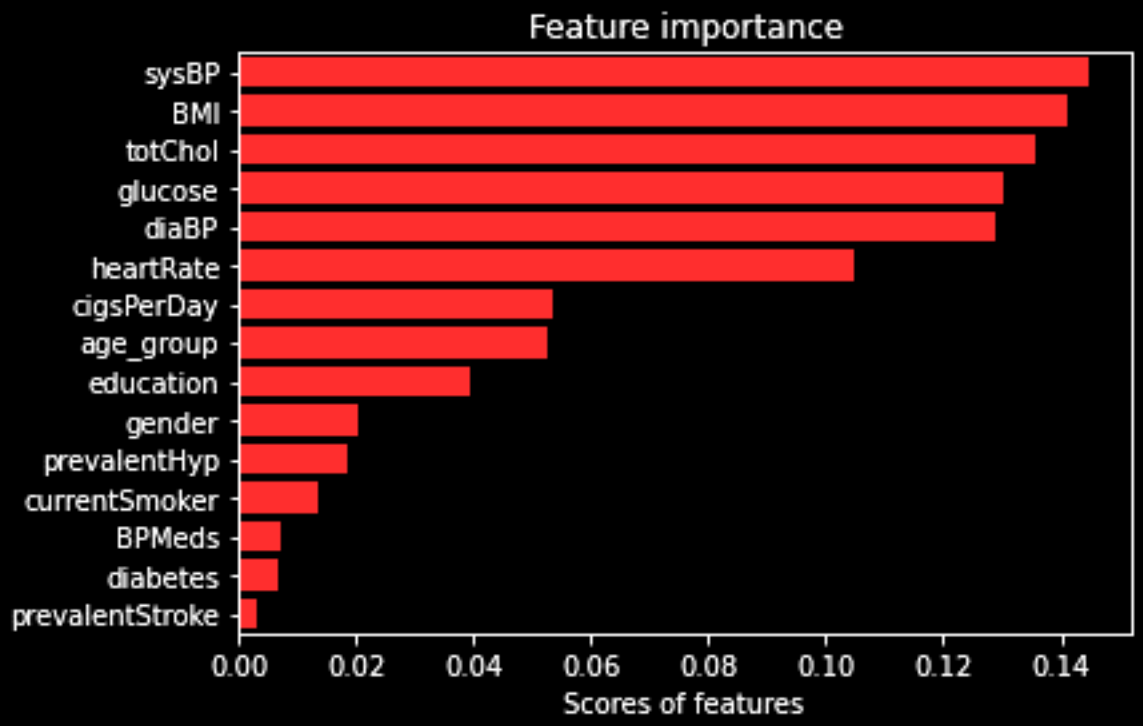
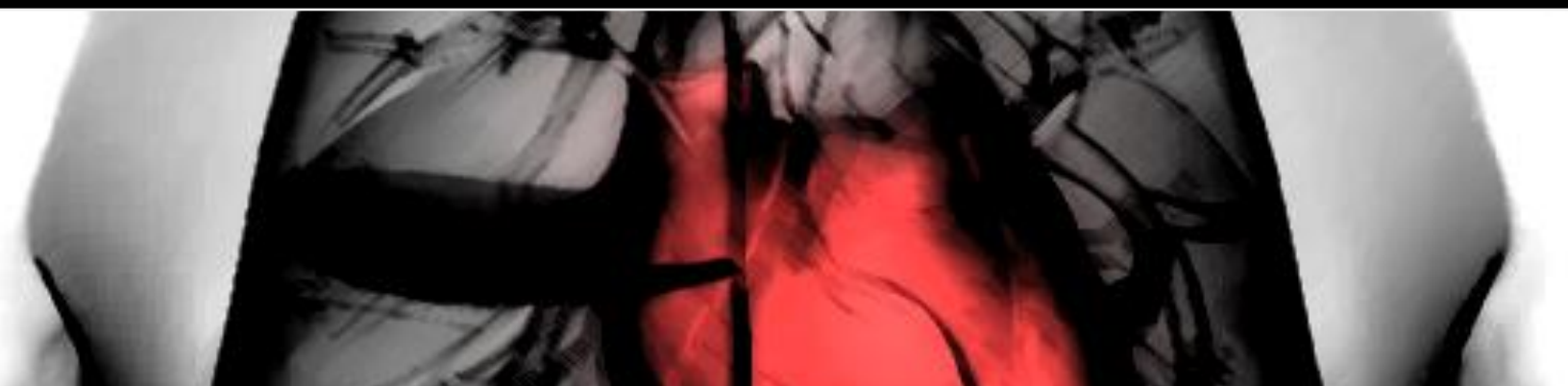


Figure 7: *Feature importance using random forest for attributes that contribute to CHD.*



Machine Learning

The goal of this step was to generate classify whether an individual is likely to suffer from coronary heart disease (CHD) or not based on demographic, comorbidity, vital sign and some laboratory investigation data. I evaluated various machine learning models with three datasets: I) using all the features, II) using six features identified as significant by logistic regression, and III) using top ten features identified by random forest model. Models performed best when the top ten features identified by random forest model were used. Therefore, I am presenting the result obtained by analyzing the top ten features. Data was split into 70% training and 30% test sets and the predictor variables were standardized using a RobustScaler. I decided to evaluate six different classifiers and select the model that performs the best. The classifiers evaluated were:

1. *KNN*
2. *Logistic Regression*
3. *Random Forest*
4. *Gradient Boost*
5. *Support Vector Machine (SVM)*
6. *Naive Bayes*

Table 2: ROC-AUC score of various algorithms

Algorithm	ROC-AUC score
Logistic Regression	0.732449
Naive Bayes	0.718285
Random Forest	0.717543
Gradient Boost	0.713462
KNN	0.696594
SVM	0.667709

Logistic regression has the highest ROC-AUC score (**Table** 2; **Figure** 7); therefore, I decided to use a logistic regression model. Classification report of the logistic regression model is presented in **Table** 3. Based on this analysis, recall to predict the occurrence of CHD is 0.06 suggesting that the model performs very poorly to correctly predict the likelihood of suffering from CHD.

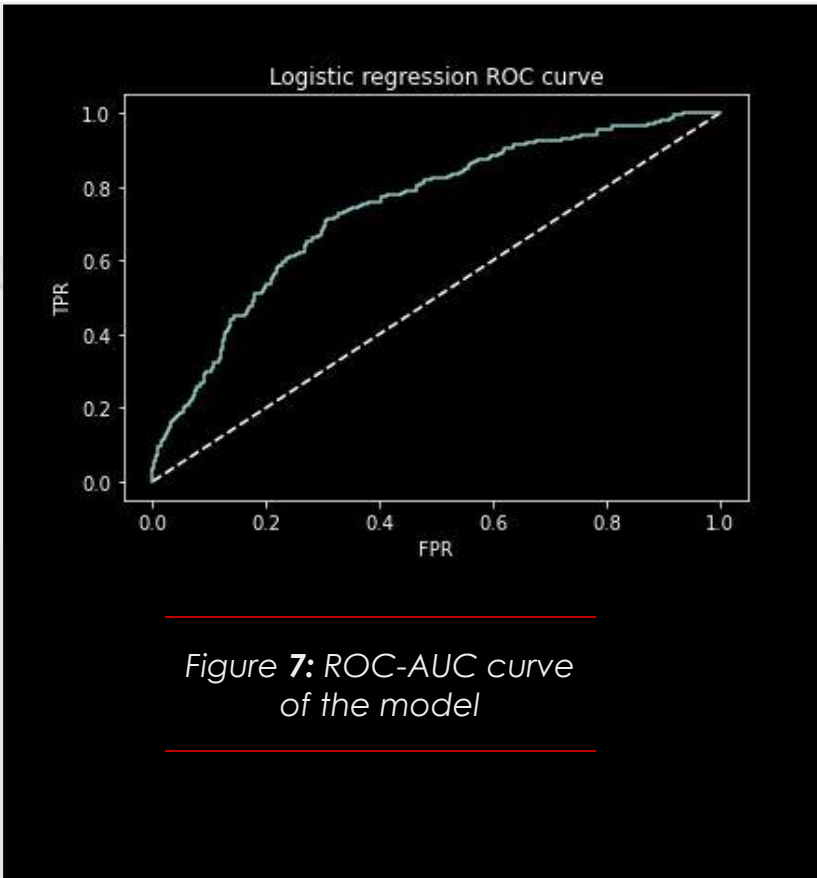


*Figure **7:** ROC-AUC curve of the model*

Adjusting the Probability Threshold

Most of the machine learning algorithms that are used for classification were formulated with the assumption of a similar number of samples across each class. If the dataset is imbalanced, the model often results in poor prediction particularly for the minority class. The dataset used in this project was also imbalanced and the model performed poorly to predict the occurrence of CHD (minority class) with logistic regression using the default threshold. One of the strategies to improve the correct prediction of minority class under such circumstances is to tune the probability threshold that assigns data to different classes such that it yields the highest F1 score. For a patient and a healthcare professional, it is always more important to correctly predict the likelihood of occurring CHD than correctly predicting the likelihood of not having CHD (ie. we want to maximize the recall value); therefore, we set the beta score of 2. When I tested across a range of threshold, I found that at a threshold of 0.168, F2 score is the highest (**Figure** 8). This result showed that when the threshold was changed from 0.5 (default) to 0.168, recall for predicting CHD increased from 0.06 to 0.73. However, the precision decreased from 0.73 to 0.30 (**Table** 3 and 4).

Table 3: Classification report of logistic regression with default threshold

	Precision	Recall	F1-Score	Support
No CHD	0.85	1.00	0.92	923
CHD	0.73	0.06	0.12	175
Accuracy			0.85	1098
Macro avg	0.79	0.53	0.52	1098
Weighted avg	0.83	0.85	0.79	1098

Table 4: Classification report of logistic regression with tuned threshold

	Precision	Recall	F1-Score	Support
No CHD	0.93	0.67	0.78	923
CHD	0.30	0.73	0.42	175
Accuracy			0.68	1098
Macro avg	0.61	0.70	0.60	1098
Weighted avg	0.83	0.68	0.72	1098

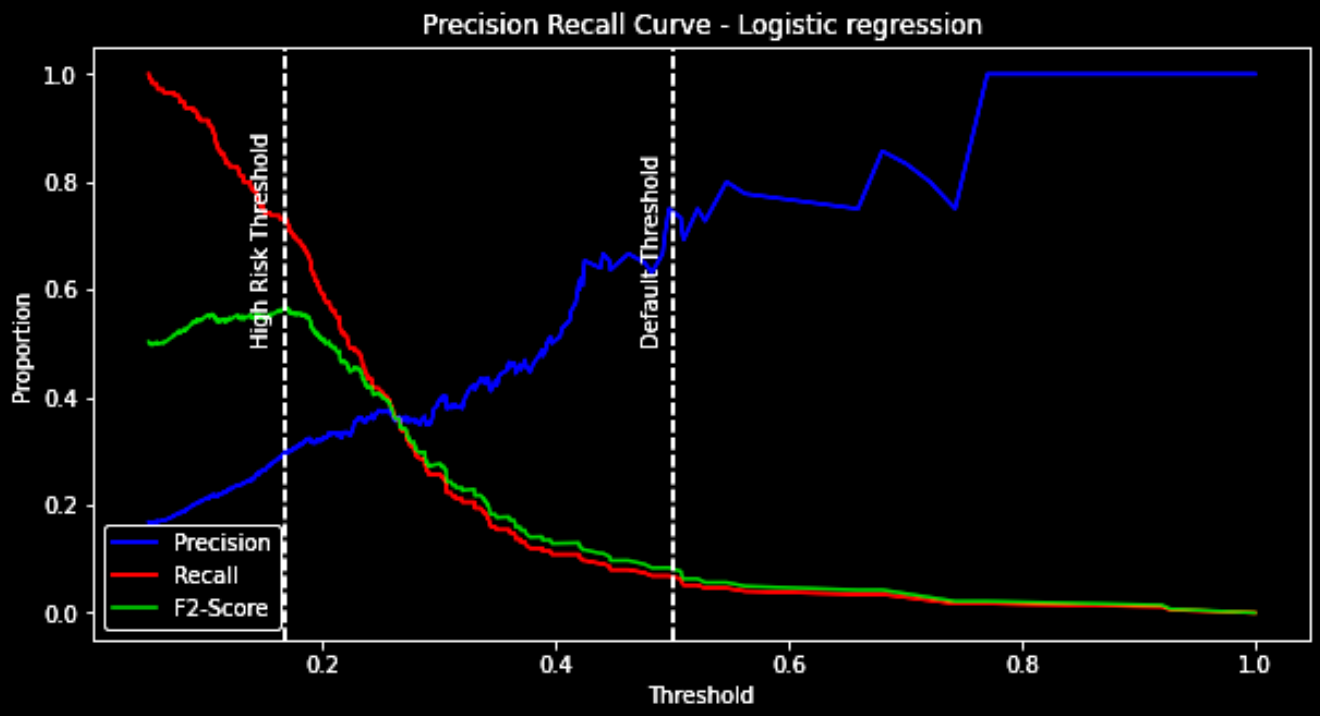


Figure 8: *Plot demonstrating F1 score, precision and recall at different thresholds*

Will undersampling or oversampling perform better?

There are other strategies to deal with an imbalanced dataset such as undersampling and oversampling. I was interested to know if undersampling or oversampling can perform better than thresholding; therefore, I evaluated various algorithms by undersampling as well as oversampling (using Synthetic Minority Over-sampling Technique (SMOTE)) such that the ratio of people with or without CHD was 1:1. I found that the logistic regression model performed best for both of these strategies. Performance of logistic regression with undersampling was similar to logistic regression after adjusting threshold with complete dataset (Recall 0.73 for predicting CHD; **Table** 5). When I tuned the threshold, the recall of the model increased to 0.79 (beta =2, threshold=0.458). The model performed the best with oversampling. Performance of logistic regression with oversampling had recall of 0.74 which improved to 0.88 after tuning the threshold (beta =2, threshold=0.392; **Table** 6). Since oversampling performed the best with the highest recall, I used this as my final model.

Table 5: Classification report of logistic regression model with under sampling (beta =2, threshold=0.458)

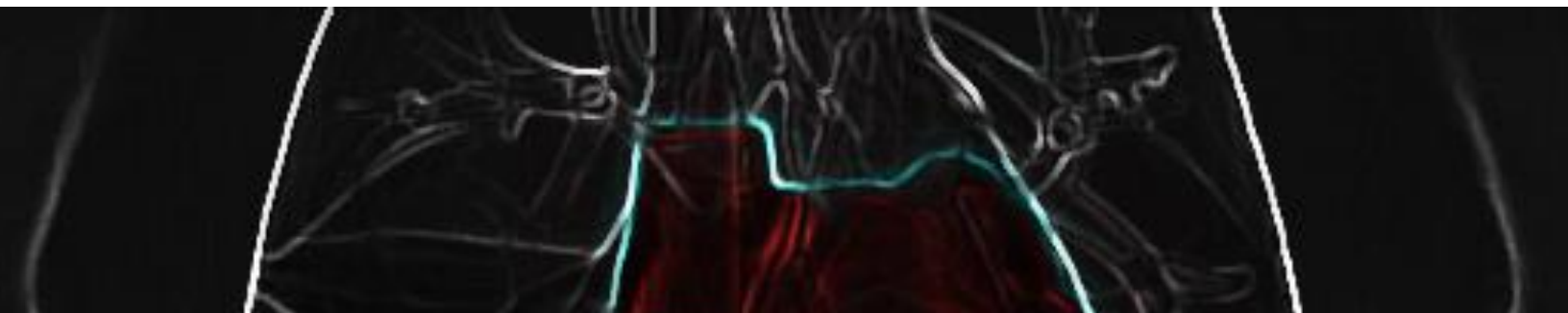
	Precision	Recall	F1-Score	Support
0	0.94	0.58	0.72	923
1	0.26	0.79	0.39	175
Accuracy			0.61	1098
Macro Avg	0.60	0.68	0.55	1098
Weighted Avg	0.83	0.61	0.66	1098

Table 6: Classification report of logistic regression model with oversampling (SMOTE; beta = 2, threshold=0.392)

	Precision	Recall	F1-Score	Support
0	0.95	0.47	0.63	916
1	0.25	0.88	0.39	182
Accuracy			0.53	1098
Macro Avg	0.60	0.67	0.51	1098
Weighted Avg	0.83	0.53	0.59	1098

Will additional data improve the model?

In the learning curve below, the training and cross-validation scores are converging together which suggests that there is enough data to make prediction and addition of more data will not improve the model significantly (**Figure** 9).



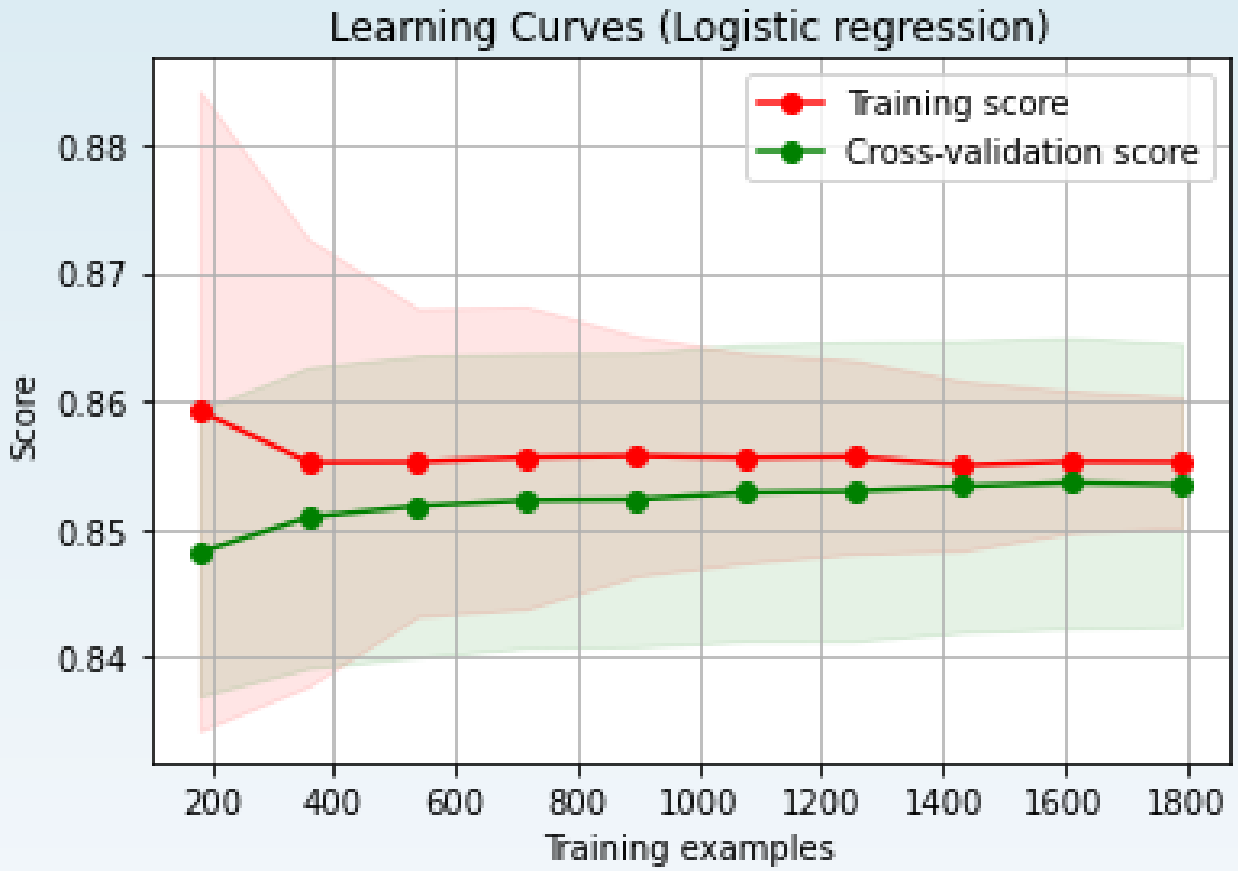


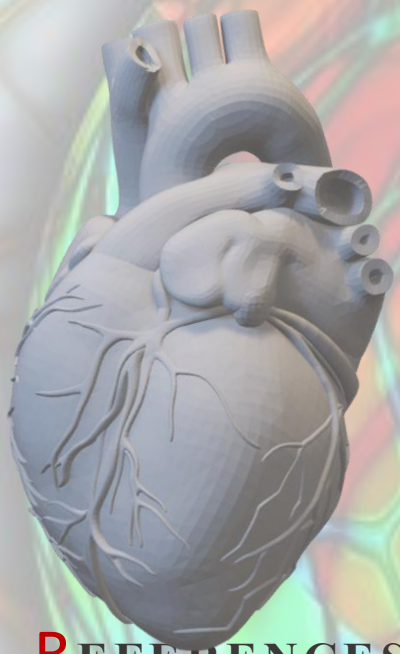
Figure 9: Learning curve of a logistic regression classifier

What opportunities exist for future improvements?

When I checked for misclassified data, I observed that there were several cases where people with hypertension that had high blood pressure, glucose level, and cholesterol had no CHD. Likewise, I also observed opposite cases. These results suggest that there are additional important features which could improve the prediction ability of the model. These features could include, for example, family history, physical activity etc. Data analyzed in this project were from the FHD original cohort study. However, FHD also includes more intensive data on second (children of original cohort) and third generation (grandchildren of original cohort), and omni cohort that reflects more diverse samples in the study. Therefore, analyzing these datasets could enhance the performance of the model.

Acknowledgement

I would like to thank Ben Bell for providing very helpful feedback to improve this project and to my family for their support.



References

CDC (2020). Heart Disease in the United States. https://www.cdc.gov/heartdisease/facts.htm. Retrieved on 1/30/2021.

World Health Organization (2017) Cardiovascular diseases (CVDs). https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Retrieved on 1/30/2021.