|  |  |
| --- | --- |
|  | Heart disease prediction dataset |

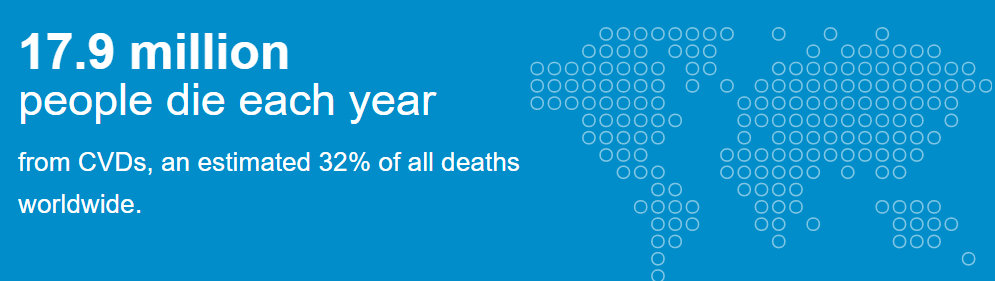
Graphical user interface, diagram

Description automatically generated



Why this Data Set:

According to World Health Organization statistics, cardiovascular diseases (CVDs) are the leading cause of death worldwide. Cardiovascular diseases are a group of diseases related to heart and blood vessels. It can be further classified into various subcategories of diseases where coronary heart disease is most common type of heart disease. The narrowing of the blood arteries resulting in stroke is known as coronary heart disease.



[[1]](#endnote-1)In 2019, an estimated 17.9 million people died from cardiovascular diseases, accounting for 32% of all global deaths. 85 percent of these deaths were caused by a heart attack or a stroke. It is defined as a medical condition which can lead to a person’s death. Since stroke has a serious impact on global health, we decided to look at the few parameters leading to it. Thus, early detection and prevention of this condition is important keeping in mind the global health conditions.

Project Overview:

During the training phase, a classification model will be constructed using several independent variables such as male, age, cigsPerDay, totChol, sysBP, and glucose, as well as a response variable (TenYearCHD class). The classification goal is to predict whether a patient has a 10-year risk of developing coronary heart disease (CHD).

Methodology:

Based on input parameters such as gender, age, education, and current health, we will attempt to predict whether a patient is likely to develop heart disease in the next ten years.

Data Overview:

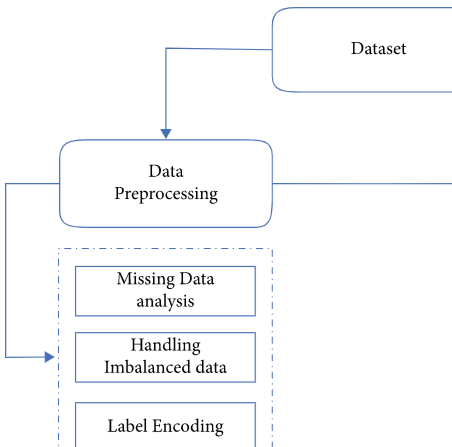
The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The analysis was performed on a heart dataset comprised of dimensions.

|  |  |
| --- | --- |
| **Rows** | **Columns** |
| 4238 | 16 |

In this section, we will predict whether a patient has a 10-year risk of developing Coronary Heart Disease (CHD). The variables with description and data type are as follows

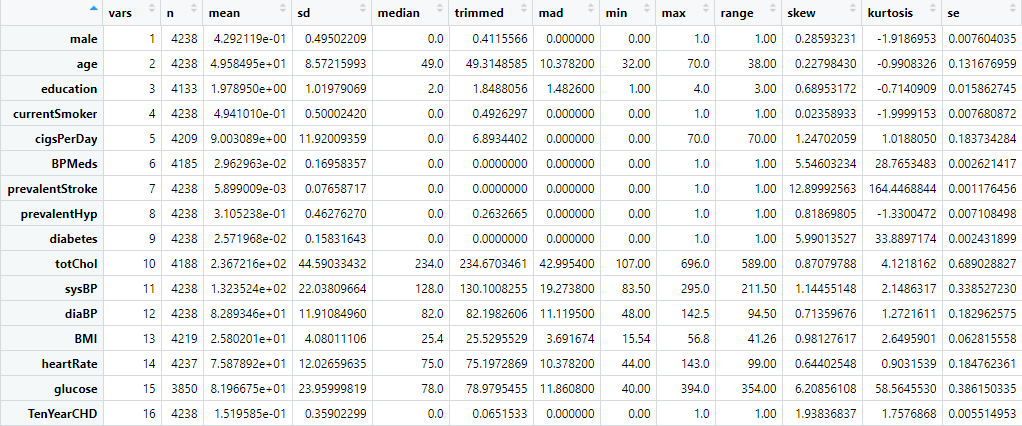
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Kind of Variable- Type** |
| **Male** | Male (1) and female (0) represent to gender | Nominal-Categorical |
| **Age** | Age of patient | Continuous-Numerical |
| **Education** | Education level of patient | Ordinal-Categorical |
| **currentSmoker** | Whether or not patient is current smoker | Nominal-Categorical |
| **cigsPerDay** | Number of cigarettes person smokes on average in one day | Continuous-Numerical |
| **BPMeds** | A patient was on blood pressure medication or no | Nominal-Categorical |
| **prevalentStroke** | A person previously had a stroke or not | Nominal-Categorical |
| **prevalentHyp** | A patient has a hypertension or not | Nominal-Categorical |
| **diabetes** | A patient has diabetes or not. | Nominal-Categorical |
| **TenYearCHD** | 10 year risk of coronary Heart disease (1: yes, 0:no) | Nominal-Categorical |
| **totchol** | total cholesterol level of patient | Continuous-Numerical |
| **sysBP** | systolic blood pressure of patient | Continuous-Numerical |
| **diaBP** | diastolic blood pressure of patient | Continuous-Numerical |
| **BMI** | Body Mass Index | Continuous-Numerical |
| **Heart Rate** | heart rate of patient | Continuous-Numerical |
| **Glucose** | Glucose level of patient | Continuous-Numerical |

Data Inspection and cleaning:



The figure illustrates the various steps in performed in data preprocessing.

* Summary statistics of the dataset is as follow:



* Here, n denotes the number of values in the dataset.
* The mean represents the average value of each column.
* The standard deviation tells us about the spread or dispersion of our data points around mean value.
* The median is the middle value of each column, and the mean is the average of each column.
* The trimmed explains the mean after trimming minimum and maximum value.
* The mean absolute deviation tells about the variation of each point from the mean value.
* The min represents the minimum value in each column.
* The max represents the maximum value in each column.
* The range represents the difference between maximum and minimum value.
* Skew represents the shape of each column's distribution, with positive values representing right-skewed distributions and negative values representing left-skewed distributions.
* Kurtosis denotes the apex of the data distribution. If the value in the column is negative, it indicates a flat data distribution, whereas a positive value indicates a peak distribution.
* Missing values:

The Missing value in dataset is as follow:





We discovered missing values for few columns. The numerical columns “glucose”, “CigsPerDay”, “totChol” and “BMI” had missing value less than 9.2% of the total values, thus, the values were imputed with the median of the non-missing values. All the other categorical columns that had missing values less than 5% so that values were dropped from the dataset to reduce biasness.

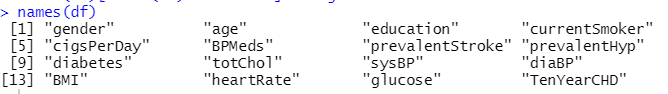
* Duplicate check:

The given dataset has no duplicate values in the dataset.



* Rename column:

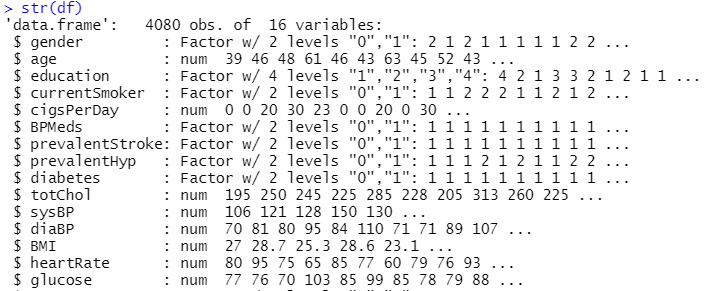
The column “male” was renamed as “gender” and the column had values “1” for “male” and “0” for “female”.



* Transforming data:

After checking the structure of the dataset, it was observed that there were categorical columns with datatype “integer” and some of the columns were of datatype “num”.

The categorical variables “gender”, “education”, “currentsmoker”, “BPMeds”, “prevalentStroke”, “prevalentHyp” and “diabetes



* Outlier Detection:

The boxplots were plotted for the numeric variables to check for any outliers in our data.

|  |  |  |
| --- | --- | --- |
| **BMI** | **AGE** | **Heart Rate** |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **Glucose** | **Diabp** | **SysBp** |
|  |  |  |

|  |  |
| --- | --- |
| **totchol** | **CigsPerDay** |
|  |  |

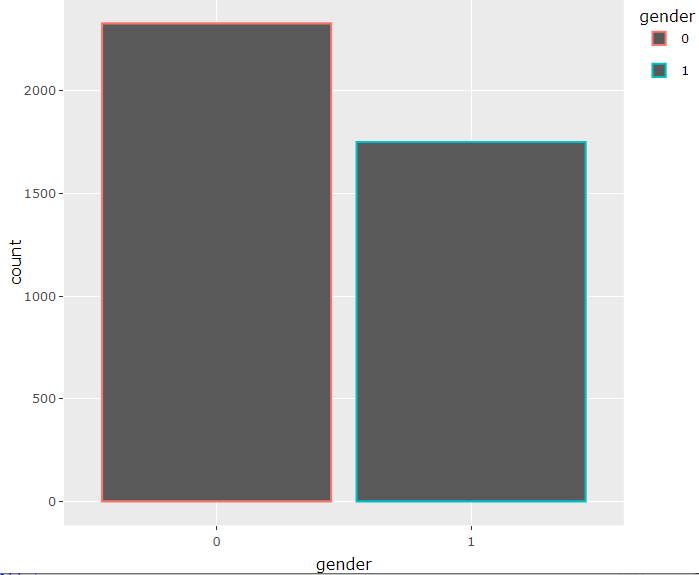
Observation:

We observed that there were a few outliers for the variables “bmi”, “heartrate”, “Glucose”, “diaBP”, “sysBP”, “totchol” and “CigsPerDay”. We decided not to drop or treat the outliers as these values were significant for further analysis.

* Distribution of Categorical Variables:

For the descriptive analysis of the categorical variables, bar graphs were plotted to check for the distribution of each categorical variable output.

|  |
| --- |
| **Gender TenYearCHD** |

 Chart, bar chart

Description automatically generated

|  |
| --- |
| **Diabetes BPMeds** |

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

|  |
| --- |
| **Education currentSmoker** |

Chart, histogram

Description automatically generated Chart, bar chart

Description automatically generated

|  |
| --- |
| **prevalentHyp prevalentStroke** |

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Observation

* The probability of having coronary heart disease in ten year is very less.
* Most of the participants were non-diabetic.
* Most of the participants were not taking BP medications.
* There were almost equal proportion of participants who were current smoker and were not.
* About 65% of participants do not have any prevalent hypertension history.
* Only a few participants had history of prevalent stroke.

## Correlations between variables:

## Numeric variables

We plotted correlation plot to find if any variables were correlated with each other.

Chart, waterfall chart

Description automatically generated

It can be observed that:

* “diaBP” is highly positively correlated with the “sysBP”.
* “diaBP” and “sysBP” are positively correlated to the “BMI”.
* “Age” is moderately positively correlated to “sysBP”.
* “Age” is less positively correlated to the “totChol”.
* “CigsPerDay” is slightly negatively correlated with all other variables except “heartrate”.

## Categorical Variables

The association between the categorical variables was checked using Chi- Square test. We tried looping the variables to check the dependency of one variable with all other categorical variables.

Text

Description automatically generated

## Initial Variable Selection

The numerical variables and categorical variables were tested to check if the variables are significant to be including in the regression model to detect the possibility of coronary heart disease in ten years.

* **Numeric variable**

For the numeric variables, we used t-test to check if the variable is statistically significant or not.

|  |  |  |
| --- | --- | --- |
| **Variable** | **t-value** | **p-value** |
| Age | -14.88 | < 2.2e-16 |
| CigsPerDay | -3.3382 | 0.0008819 |
| totChol | -4.5989 | 4.932e-06 |
| sysBP | -11.838 | < 2.2e-16 |
| DiaBP | -8.1974 | 1.037e-15 |
| BMI | -4.3731 | 1.387e-05 |
| heartrate | -1.3526 | 0.1825 |
| glucose | -4.6366 | 4.268e-06 |

The p-value is less than all other variables except “heartrate”.

**The hypothesis for t-test is:**

**Null hypothesis:** The mean group difference is zero.

**Alternate hypothesis:** The mean group difference is not zero.

We can conclude that “heartrate” is not significant for any further analysis.

* **Categorical variables**

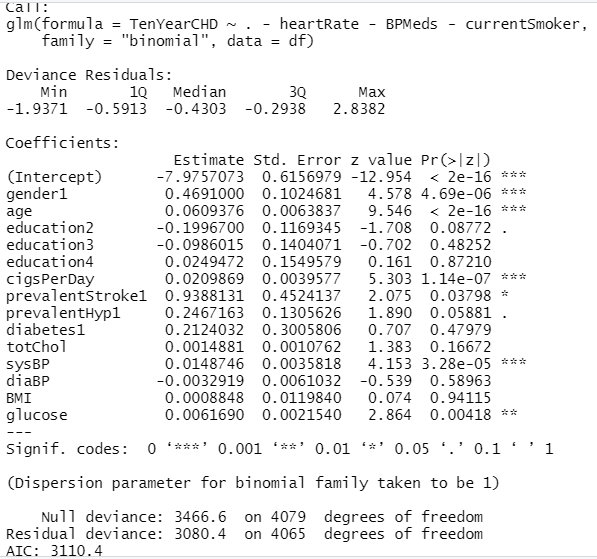
For the categorical variables, we used **Chi-Square test** to check if the variable is statistically significant or not. We ran a loop for testing each variable association with all other variables.

A close-up of a document

Description automatically generated with low confidence

## Building Logistic Regression

We ran a logistic regression model considering all the variables as explanatory variables except “heartrate” and “current smoker”.

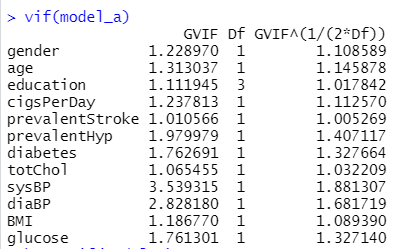


We can conclude from the output of the model that

* “gender”, “age” and “cigsPerDay” are highly significant in predicting the probability of having coronary heart disease in ten years.
* “sysBp” is moderately significant in the model.
* “prevalentHyp”, “totChol” and “glucose” are less significantly associated in predicting CHD.

## Checking Multicollinearity

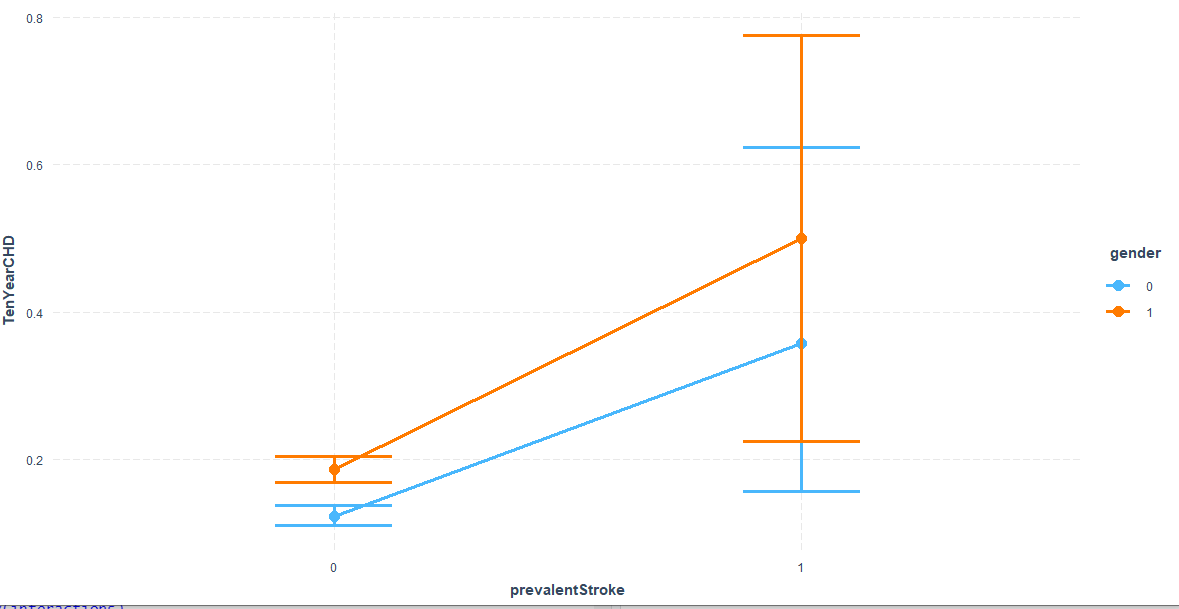
The model was then tested for any multicollinearity issue and Variance Inflation Factor (VIF) was used.



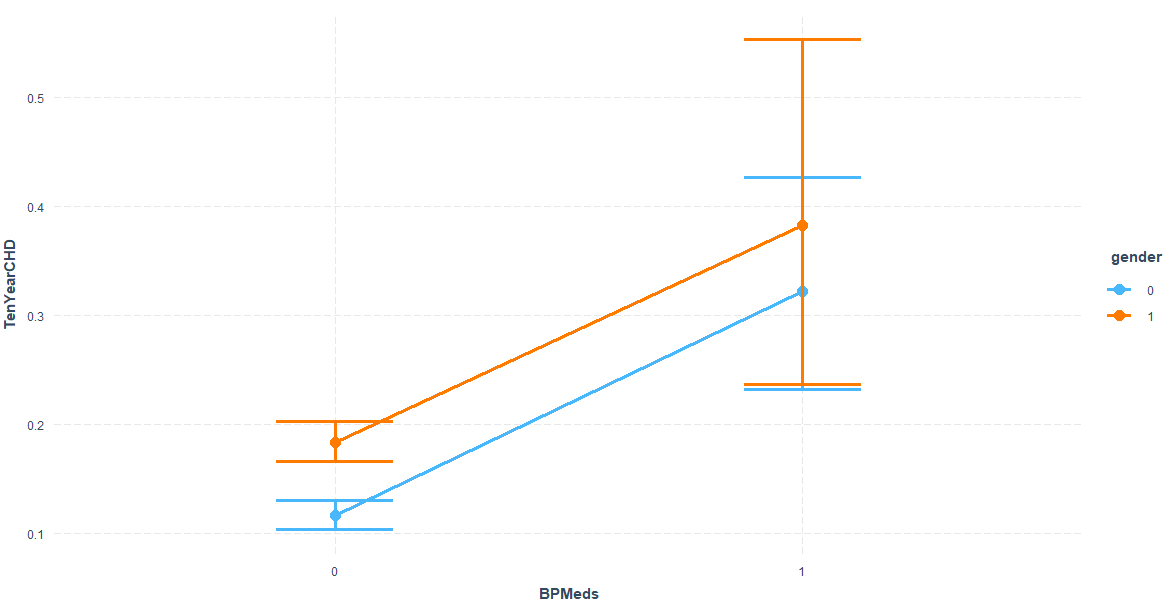
No vif value is greater than 5, so there is no multicollinearity in our model.

## Interaction Analysis

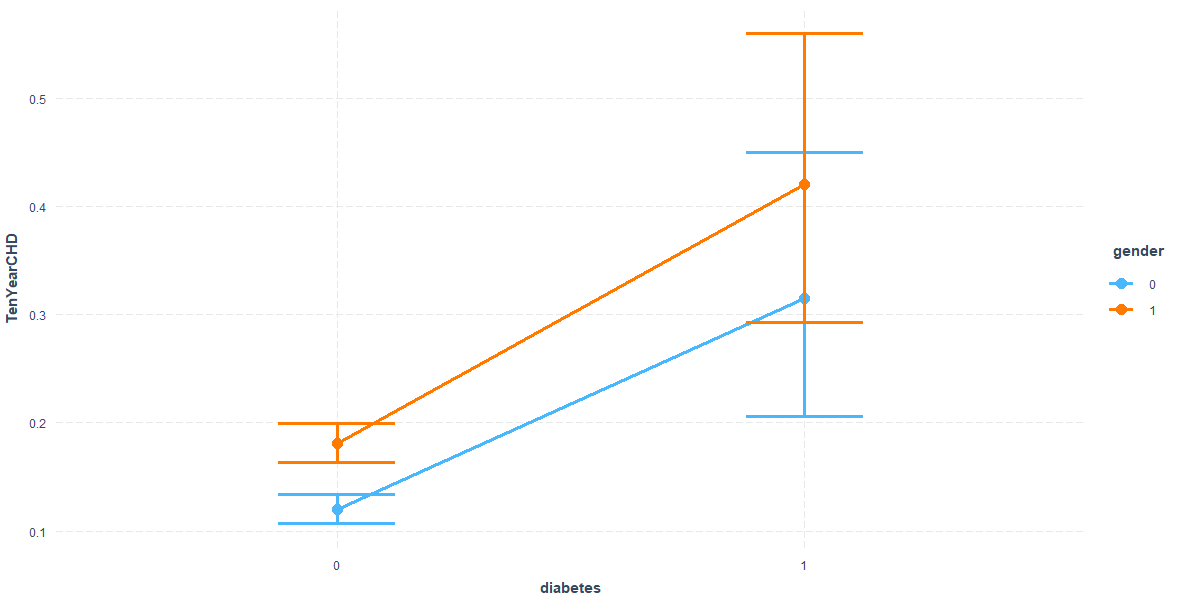
We analyzed various interaction plots, Here are some



As, we have observed from the interaction plot between Prevalent Stroke , TenYearCHD and Gender that there is interaction between these variables as the lines are not parallel and will intersect at one point to the left of the plot



In the interaction plot of TenYearCHD, BPMeds and gender we see parallel lines which says that there is no interaction between these variable.



Observation:

In the interaction plot between TenYearCHD, diabetes and gender. We see that there is interaction between these variables as the lines are not parallel and will intersect at one point to the left of the plot

**Data Splitting**

For Data Splitting used 80-20% splitting technique.

|  |  |  |
| --- | --- | --- |
|  | **Obsevation** | **Variables** |
| **Training Data** | 3264 | 16 |
| **Testing Data** | 816 | 16 |

**Remark:** Used splitting technique to check the Accuracy of the model.

**Variable Selection Technique**

We used backward selection technique

* Backward selection technique starts with all variables and removes those which are not significant to the model.
* At the end the leave only those variables that contribute substantially to the model.

|  |  |
| --- | --- |
| **Without Interaction** | **With Interaction** |
|  |  |
|  |  |

It has been observed that the interaction plot has some extra variables. Moreover, we can see that the final model has AIC 2508.76 in without interaction term and with interaction terms has AIC 2510.66. which shows that Without interaction performs better which shows that interaction terms are not significant our further analysis

**Model without Interaction Terms:**

TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke +

sysBP + glucose

**Model with Interaction Terms:**

TenYearCHD ~ gender + age + cigsPerDay + prevalentStroke + prevalentHyp +

totChol + sysBP + diaBP + glucose + sysBP:diaBP

**Model Comparison**

For model Comparison we can Check AIC and test using Log Likelihood Ratio Test. The result is as given bellow:

|  |  |
| --- | --- |
| **Without Interaction** | **With Interaction** |
|  |  |

**AIC**

|  |  |  |
| --- | --- | --- |
|  | **Without Interaction Term** | **With Interaction Term** |
| **AIC Value** | **2508.87** | **2510.7** |

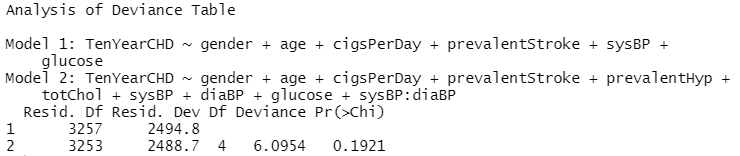
AIC is Akaike information criterion. Lower the AIC the better the model is. Hence it is cleared that model without interaction term has lower AIC value than With Interaction Term.

**Log Likelihood Ratio Test**

Log Likelihood is used to check which model is better in terms of significance. In this we have taken Model with interaction and Model without interaction terms.

H0: Reduced model is appropriate

Ha: Full Model is appropriate



It shows that p-value is greater than 0.05 which means we accept null hypothesis says model without interaction terms is better.

**Prediction Accuracy**

To check the Accuracy of the data we checked classification report and ROC curve for both with and without interaction model. For this model is trained on training dataset and to check accuracy we used testing dataset.

**Classification Report**

It is the detailed table which we got from R console which tells us about the Accuracy, sensitivity and specificity of the model

**Model without Interaction**

|  |  |
| --- | --- |
|  |  |

* The accuracy rate of the model is 85.66%. Whereas Misclassification rate is 14%
* The Sensitivity of detecting positive case is 99.7% that is people likely to have heart disease in 10 years and is predicted by the model.
* The specificity of detecting the negative case is 0.05

**Model With Interaction**

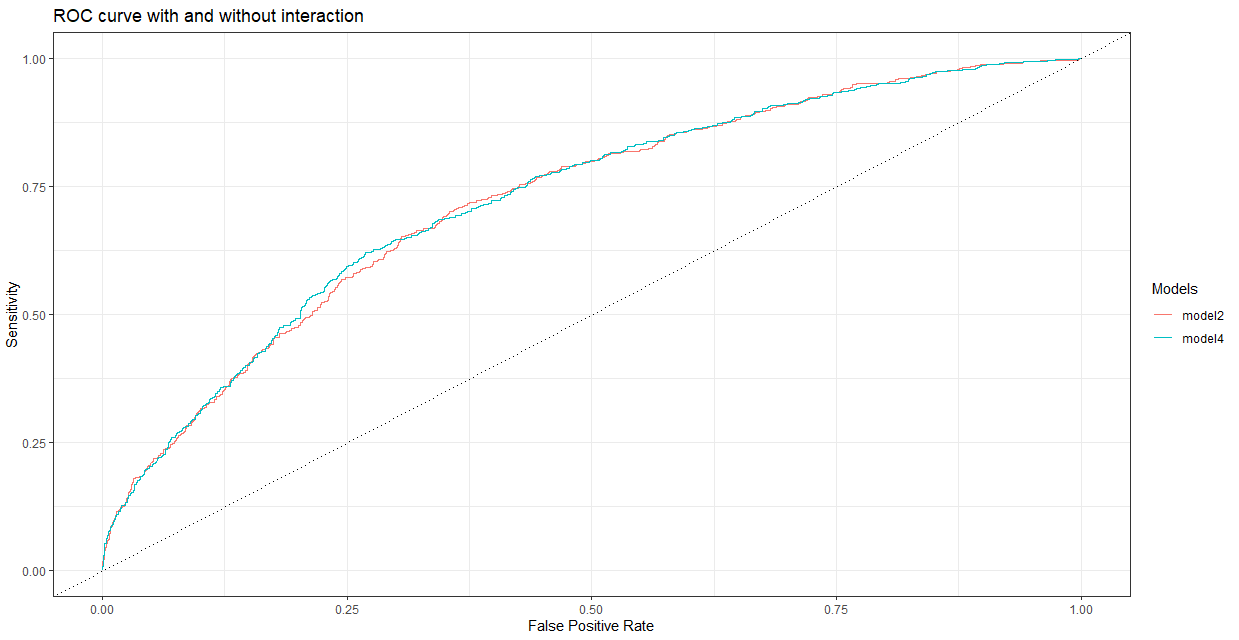
|  |  |
| --- | --- |
|  |  |

* The accuracy rate of the model is 85.54%.
* The Sensitivity of detecting positive case is 99.5% that is people likely to have heart disease in 10 years and is predicted by the model.
* The specificity of detecting the negative case is 0.05.

**ROC Curve**

ROC curve is the probability curve. It indicates how well the model can distinguish between classes. Where AUC measures the separability. Hence, Higher the AUC the better the model is for prediction

The ROC curve with and without interaction in one plot. As shown below



**Observation:**

Model2 – AUC value = 0.722

Model4- AUC Value = 0.724

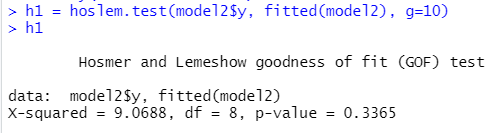
From, plot it has been observed that AUC value is approximately same for model which is 72%. So, we consider our model without interaction plot is better as it will be less complex in analysis.

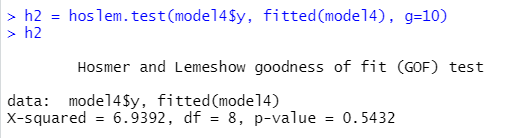
**Hosmer-Lemeshow Test**

Hosmer Lemeshow test is a goodness of fit test used in logistic regression. It tells us how the data fits the model.

H0: Reduced model is appropriate

Ha: Full model is appropriate





The p-value of both model is greater than 0.05, so we fail to reject null hypothesis which shows that our model is good fit to predict whether a patient has a 10-year risk of developing coronary heart disease (CHD).

**Conclusion:**

Coronary heart disease is a life-threatening disease which should be treated at the earliest. So, in our project report, we analyzed some of the factors that might cause it. As they are impacting global health therefore doing early detection and appropriate management can prevent any further loss. We have analyzed the various risk factors responsible for a person to develop CHD (coronary heart disease) in ten years thereby prevention of these can help in reducing the risk of developing the heart disease. The factors that influence if a person will have a coronary disease in 10 years are gender”, “age” and “cigsPerDay. We got three models one with backward selection, one with interaction terms and one without interaction terms but after comparison we have chosen the model without any interaction terms as it is less complex and there is data imbalance in our model so the accuracy of both models is almost identical with a slight variation. The high sensitivity indicates that our model will correctly predict if a person will have coronary disease in 10 years and less specificity indicating that the model will not correctly predict if the person will not develop CHD in 10 years.

The insights found in the dataset are:

* “gender”, “age” and “cigsPerDay” are highly significant in predicting the probability of having coronary heart disease in ten years.
* “sysBp” is moderately significant in the model.
* “prevalentHyp”, “totChol” and “glucose” are less significantly associated in predicting CHD.
* The model without the interaction terms is better and less complex than the model with the interaction terms.
* Our model has accuracy rate of 86%.
* Our model is more sensitive than specific.
* There is a presence of data imbalance in our model.

**Recommendation:**

* The methodology used for the data collection could have been improvised by selecting more significant factors contributing if a person will develop coronary heart disease such as being obese, body fat percent, stress and having a family history of CHD.

**References:**

1. WHO. (2021). Cardiovascular diseases. *Health topics*. [↑](#endnote-ref-1)