



Technical Report for Advance Data and Networking Mining CA2

By-

Muktesh Sahu (10594774)

Maya Kumbhar (10594952)

Ankit Butola (10590492)

Contents

Abstract	3
Introduction.....	3
Key Business Stakeholder	4
Needs of the Business	4
Analytical Solutions.....	4
Business Profits and Challenges.....	5
Reporting and the Visualizations.....	5
Aspects of the Analytical-Issue	5
ABT - Attributes and Results	5
Hypotheses	6
Key for the Success of the Metrics.....	6
Data and its sources	6
The Requirements of Data and its Sources	6
Quality and Visual Exploration	7
Data Harmonization, Rescale, and Cleaning	8
Refining the problem definition	9
The Methodology Technique	10
Methods to solve the potential problems.....	10
Tools/Software Required	10
Testing and Model Selection Criteria	11
Model Development	15
Model Structure Identification	15
Running and Evaluating the Model	16
Optimization and Recalibration of the Model	17
Threshold Technique & Result.....	17
Smote Up sampling and its Outcome	18
Documentation of Research results	19
Assumptions on the result	19
Limitations of the approach.....	19
Constraints.....	19
Deployment	20
Model Business Validation.....	20
Conclusion and Recommendations	20

Abstract

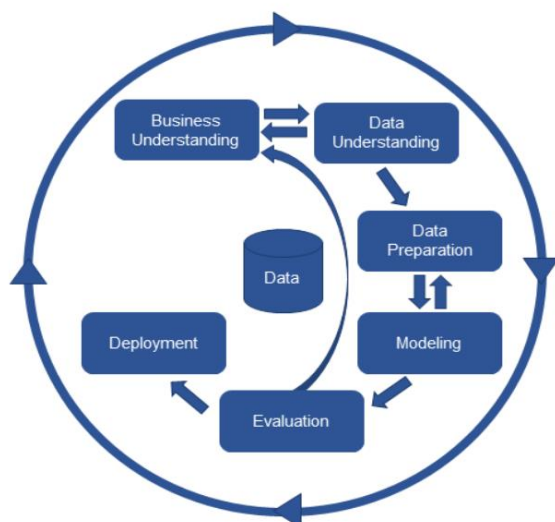
Cases of cardiovascular disease (CHD) and variables associated with the condition of the patient are included in the 10YEARS_CHDdata.csv data set: systolic pressure in the blood, consumption of tobacco annually by an individual (in kg), low- lipoprotein density check (ldl), consumption of alcohol by individual, their age, and the main measure which is the C.H.D. prediction among the population (0 or 1). In this research, we discuss the strategies used by the Machine Learning which helps to implement the model which in turn helps to predict the risk/ the chances over C.H.D after 10 years.

The presented research technique is separated into four segments: data capture, pre-processing and then moulded to obtained data to the preferred format of business needs, designing the intelligent models employing machine learning technique, splitting the C.H.D data in to the training and testing sets, and lastly, we performed the validation of the model using stratified sampling technique.

At the end, we found that the regression which was performed to conduct the diagnosis using logistics regression is positive amongst the others, which thus is the preferred way when it comes to make such predictions in an accurate way.

Introduction

Data is derived from an ongoing cardiovascular study of residents of Farmingham, Massachusetts, United States of America, with the goal of detecting a 10-year risk of developing coronary heart disease in the context of all limitations.



This is a serious disease that is quite common among all of the world's prominent diseases. Annually, approximately 12 million people die as a result of heart diseases worldwide.

Hearing illness claims the lives of 6,10,000 people in the United States alone each year.

This is a frightening fact: that in U.S.A. CHD is the most leading reason of causing death in both the genders male/female.

To be precise, CHD can be stated as the common cause which end up killing annually 370,000 individuals.

Key Business Stakeholder



Some of the most common stakeholders for our project are,

- Health care: The health-care research industry has a strong focus on the heart division.
- Insurance segment: Insurance firms must detect risks among their customers.
- R&D: Industries involved in medical research.

Needs of the Business

Build a Ten Years CHD for each patient relating to the following criteria:

- Basis on level of educational qualifications, a individual's risk of having coronary heart disease (C.H.D).
- With the prevalence of hypertension attribute, an individual's risks of to have this disease.
- With blood pressure, we can detect a individual's risks of C.H.D.
- The likelihood of a somebody developing coronary heart disease is determined by their gender.
- Based on the prevalence of diabetes, a person's possibilities of developing coronary heart disease.
- The likelihood of an individual developing coronary heart disease is determined by their smoking behaviours.
- Based on the prevalence of stroke, a patient's chances of developing C.H.D.

Analytical Solutions



There are many different derived random secret patterns and the considerations over it with which we can foresee and make the future prediction and the main task of this activity is to detect and foresee the widespread of this disease worldwide.

If we relate it with the insurance firm than there it can help the industry to formulize the method and calculate the risk which is involved with the customers having premium subscription. Similarly, in healthcare industries

it can help in detecting and fixing challenges among future people with the condition.

Business Profits and Challenges



- The minimal amount of data accessible for analysis is a major constraint in this project.
- There are just about 4000 rows accessible.
- There are 3600 rows of data left after data cleansing.
- It's difficult to find additional facts like this one.

The capacity to foresee the onset of coronary heart disease in a patient will aid doctors in predicting and caring for such individuals. The data also could be used by the health insurance provider to determine the charge for the insurance premium subscription. Not only this, it will help the researchers in many ways while making attempts.

Reporting and the Visualizations

Aspects of the Analytical-Issue

The difficulty in predicting and then diagnosing the C.H.D. after a decade has an influence on the health sector, the treatments to be provided, and other factors that affect the individual, so studying the data and understanding what it says is a perfect way to commence.

ABT - Attributes and Results

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
Male	Age	Education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totalChol	sysBP	diastBP	BMI	heartRate	glucose	TenYearCHD

Hypotheses

- In terms of gender, 0 denotes a female and 1 denotes a male.
- Education, grouped in team of 1, 2, 3 and 4.
- To identify smokers, 0 denotes people who do not smoke and 1 denotes people who smokes.
- In case of hypertension, 0 indicates non-prevalence whereas 1 indicated prevalence to hyper tension.
- For High-pressure, 0 stands for normal BP and 1 stands for people with high BP.
- In case of diabetes, 0 refers to non-prevalence and 1 refers to prevalence.
- Similarly, next 10 years CHD (Coronary.heart.disease) is classified as 0 and 1 representing, non-prevalence and prevalence individual's.

Key for the Success of the Metrics

This project is evaluated using metrics such as,

- How reliable the outcomes are anticipated depending on the attributes used.
- How accurate is the forecasting of the outcome?
- The class's sensitivity
- The Class's Specificity
- The Classification of Error

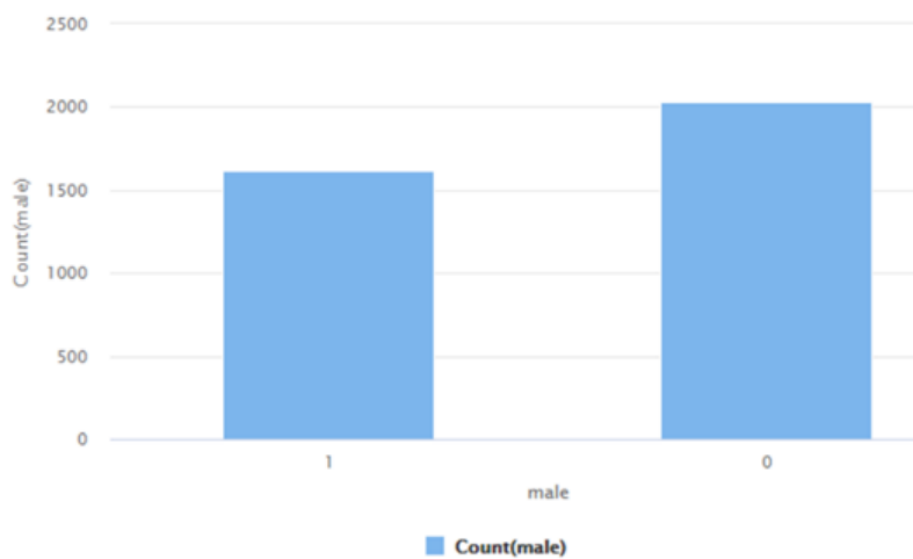
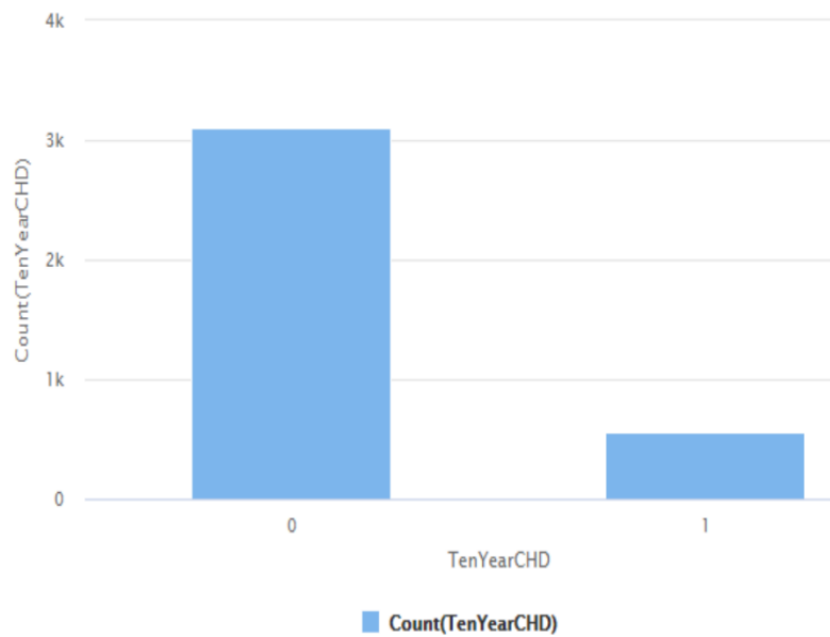
Attribute	Weight
prevalentStroke	0.443
BPMeds	0.220
male	0.179
prevalentHyp	0.078
education	0.072
sysBP	0.071
glucose	0.067
cigsPerDay	0.063

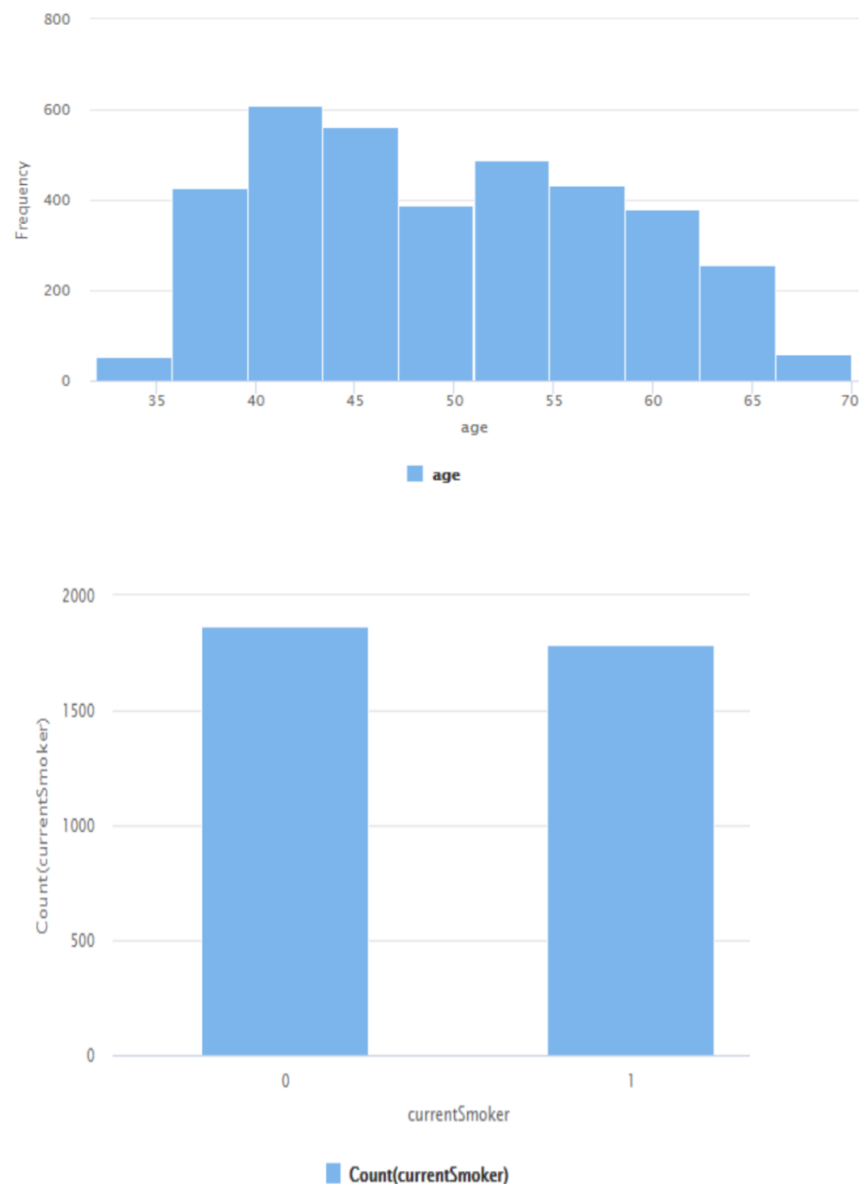
Data and its sources

The Requirements of Data and its Sources

As mentioned at the initial, the information requires specific characteristics in order to identify the 10YearCHD. The dataset's information source is the national health official online platform in the U.S.A.

Quality and Visual Exploration





Data Harmonization, Rescale, and Cleaning

We firstly loaded the data using python code in colab, we performed data pre-processing step in which we cleaned the data by eliminating the rows with no values, below are the snapshot representing the steps performed on the dataset.


```

#Importing the required Library
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import fetch_openml

[3] #Reading the dataset
framingham_heart_disease = pd.read_csv("sample_data/framingham_heart_disease.csv")

[4] #Reading few data from the dataset
framingham_heart_disease.head()

```

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	1	39	4	0	0	0	0	0	0	195	106.0	70.0	26.97	80	77
1	0	46	2	0	0	0	0	0	0	250	121.0	81.0	28.73	95	76
2	1	48	1	1	20	0	0	0	0	245	127.5	80.0	25.34	75	70
3	0	61	3	1	30	0	0	1	0	225	150.0	95.0	28.58	65	103
4	0	46	3	1	23	0	0	0	0	285	130.0	84.0	23.10	85	85

The missing values in the rows were removed rather than replaced with some parameter estimates on median because this is a medical project and abstract values can sometimes deter the consequence.

```

[10] #Data Cleaning
#Dropping the row with null values
framingham_heart_disease_New = framingham_heart_disease.dropna()

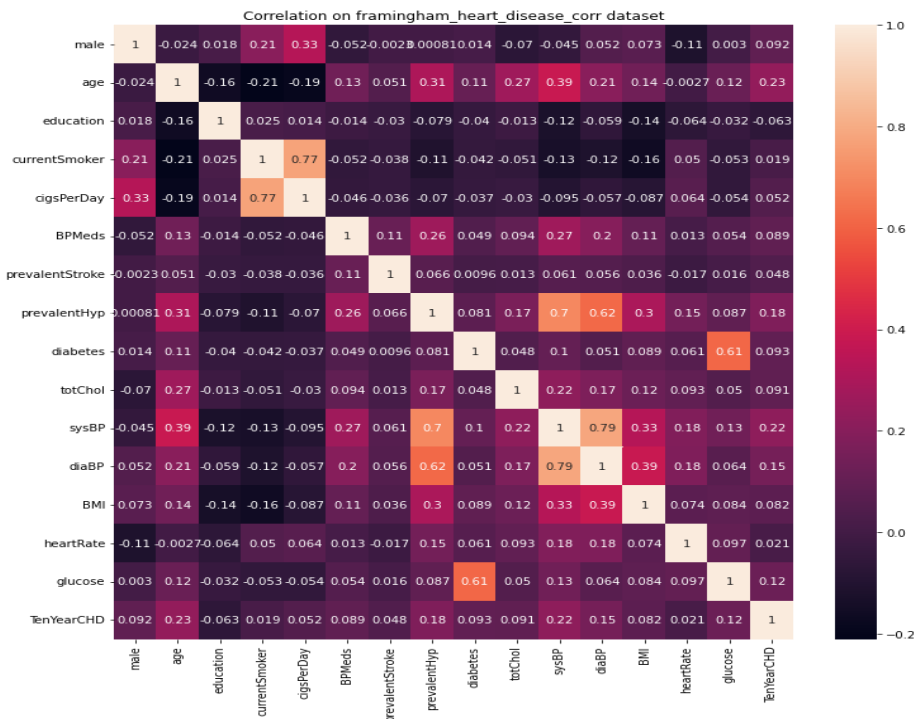
#Records with non null values
framingham_heart_disease_New.notnull().count()

```

male	3656
age	3656
education	3656
currentSmoker	3656
cigsPerDay	3656
BPMeds	3656
prevalentStroke	3656
prevalentHyp	3656
diabetes	3656
totChol	3656
sysBP	3656
diaBP	3656
BMI	3656
heartRate	3656
glucose	3656
TenYearCHD	3656
dtype: int64	

Refining the problem definition

With the help of below box plot diagram, we now have all the attributes contained in the dataset along with its defined weights. With the help of this, we can easily select those attributes which weights high than other attributes and with a lower co-relating data.



Furthermore, as we are looking after the problem based out of classification, we have investigated the variables representing categorical data and at the same time we will be rejected the variables which are grouped as continuous data having lower weight as compared to others.

The Methodology Technique

In order to resolve the problem associated with the statement, multiple methods can be carried out, we have showcased some of them through our project.

Methods to solve the potential problems

- SVM
- Logistic Regression
- Decision Tree
- Random Forest
- K Nearest Neighbor
- Naï ve Bayes
- Gradient Descent

Tools/Software Required

- RapidMiner
- Python
- Google Colab

Testing and Model Selection Criteria

We have used the auto model to automatically create and test our dataset with different intelligent models which includes Decision tree, logistic regression, random forest, generalized linear model, gradient boosted trees, deep learning, fast large margins and support vector machine.

Below table shows the measures/results that have been generated from AutoModel:

Table 1. Logistic Regression

Criterion	Value	Standard Deviation	
Logistic Regression			
accuracy	0.8354067	0.020745917	
classification_error	0.1645933	0.020745917	
AUC	0.64784981	0.027221277	
precision	0.44	0.334290293	
recall	0.0426126	0.015313382	
f_measure	0.07460358	0.024855066	
sensitivity	0.0426126	0.015313382	
specificity	0.98294674	0.012761046	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	866	157	0.8465
pred. 1	15	7	0.3182
class recall	0.983	0.0427	

Table 2. Naïve Bayes

Criterion	Value	Standard Deviation	
Naïve Bayes			
accuracy	0.84401914	0.016780553	
classification_error	0.15598086	0.016780553	
AUC	0.66881828	0.029626839	
precision	NaN	NaN	
recall	0	0	
f_measure	NaN	NaN	
sensitivity	0	0	
specificity	1	0	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	882	163	0.844
pred. 1	0	0	0
class recall	1	0	

Table 3. Decision Tree

Criterion	Value	Standard Deviation	
Decision Tree			
accuracy	0.84401914	0.016780553	
classification_error	0.15598086	0.016780553	
AUC	0.5	0	
precision	NaN	NaN	
recall	0	0	
f_measure	NaN	NaN	
sensitivity	0	0	
specificity	1	0	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	882	163	0.844
pred. 1	0	0	0
class recall	1	0	

Table 4: Generalized Linear Model

Criterion	Value	Standard Deviation	
Generalized Linear Model			
accuracy	0.8354067	0.020745917	
classification_error	0.1645933	0.020745917	
AUC	0.64784981	0.027221277	
precision	0.44	0.334290293	
recall	0.0426126	0.015313382	
f_measure	0.07460358	0.024855066	
sensitivity	0.0426126	0.015313382	
specificity	0.98294674	0.012761046	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	866	157	0.8465
pred. 1	15	7	0.3182
class recall	0.983	0.0427	

Table 5: Fast Large Margins

Criterion	Value	Standard Deviation	
Fast Large Margins			

accuracy	0.83444976	0.019901085	
classification_error	0.16555024	0.019901085	
AUC	0.65138428	0.025294274	
precision	0.34	0.147478812	
recall	0.0426126	0.015313382	
f_measure	0.07424708	0.02511743	
sensitivity	0.0426126	0.015313382	
specificity	0.9818168	0.011002505	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	865	157	0.8464
pred. 1	16	7	0.3043
class recall	0.9818	0.0427	

Table 6: Deep Learning

Criterion	Value	Standard Deviation	
Deep Learning			
accuracy	0.84210526	0.015130515	
classification_error	0.15789474	0.015130515	
AUC	0.64086985	0.042804174	
precision	NaN	NaN	
recall	0.01867725	0.017431835	
f_measure	NaN	NaN	
sensitivity	0.01867725	0.017431835	
specificity	0.99436698	0.00562372	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	877	160	0.8457
pred. 1	5	3	0.375
class recall	0.9943	0.0184	

Table 7: Gradient Boosted Trees

Criterion	Value	Standard Deviation	
Gradient Boosted Trees			
accuracy	0.84401914	0.016780553	
classification_error	0.15598086	0.016780553	
AUC	0.66860265	0.01808402	
precision	NaN	NaN	
recall	0	0	

f_measure	NaN	NaN	
sensitivity	0	0	
specificity	1	0	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	882	163	0.844
pred. 1	0	0	0
class recall	1	0	

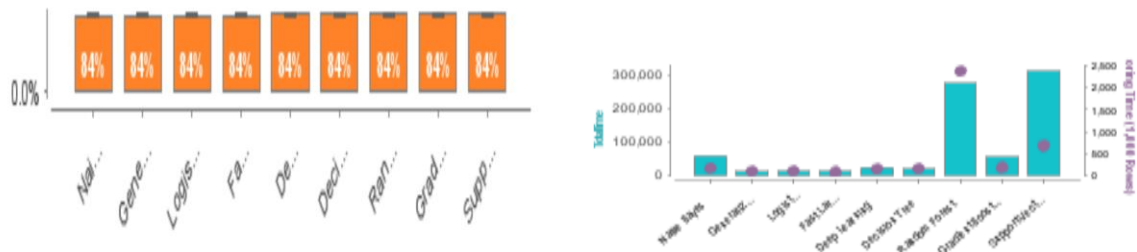
Table 8: Random Forest

Criterion	Value	Standard Deviation	
Random Forest			
accuracy	0.83062201	0.012927518	
classification_error	0.16937799	0.012927518	
AUC	0.65555934	0.027299216	
precision	0.30777778	0.071535194	
recall	0.06685398	0.030710003	
f_measure	0.10800636	0.044392817	
sensitivity	0.06685398	0.030710003	
specificity	0.97282469	0.00895125	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	857	153	0.8485
pred. 1	24	11	0.3143
class recall	0.9728	0.0671	

Table 9: Support Vector machine

Criterion	Value	Standard Deviation	
Support Vector Machine			
accuracy	0.84401914	0.016780553	
classification_error	0.15598086	0.016780553	
AUC	0.48641167	0.031064616	
precision	NaN	NaN	
recall	0	0	
f_measure	NaN	NaN	
sensitivity	0	0	
specificity	1	0	
Confusion Matrix	TRUE 0	TRUE 1	class precision
pred. 0	882	163	84.40%

pred. 1	0	0	0.00%
class recall	100.00%	0.00%	



Taking mostly three parameters of measures into consideration i.e., precision, classification error and accuracy (Run time in some cases), we have decided to go for **logistic regression** for creating the model with respect to our CHD dataset.

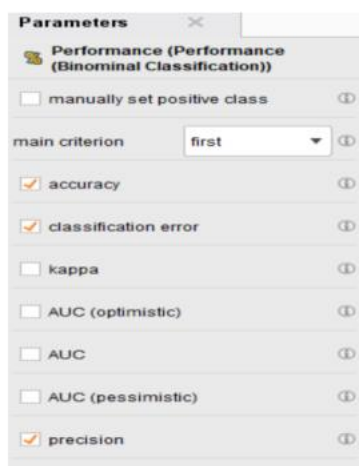
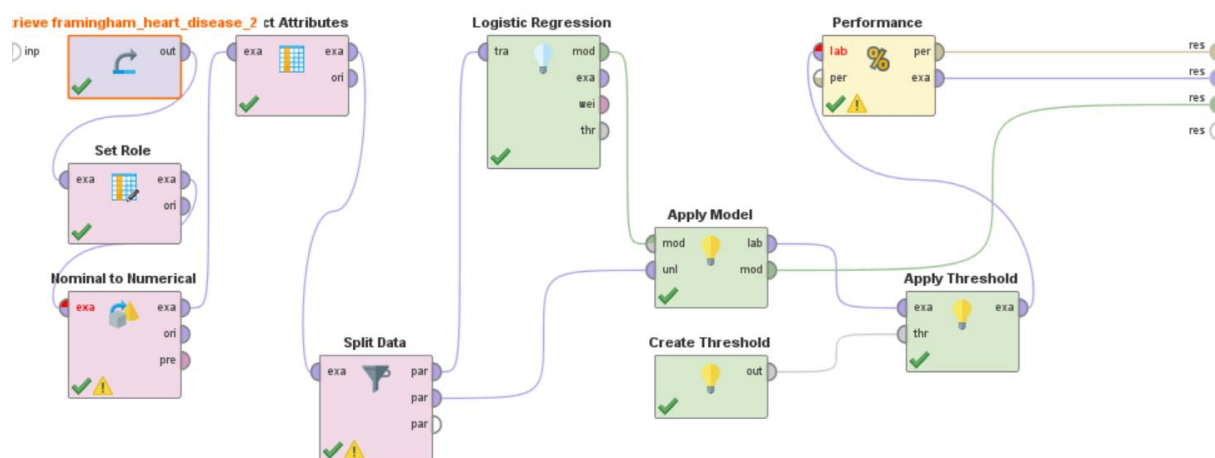
Although, automodel is considering Naive Bayes as the best performance and best gain followed by Logistic regression but Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case. Hence, we have decided to proceed with the **logistic regression** that learns the probability of a sample belonging to a certain class. We have used SMOTE for sampling and Thresholds as part of Optimization which gives variation in the accuracy as shown in the optimization section.

Model Development

Model Structure Identification

In order to build the model, we have performed the following steps:

- Choosing categorical attributes to feed into a logistic regression model in order to predict the outcome.
- Splitting the 10YRSCHD dataset into 70:30 ratio.
- To predict the values, we apply the logistic regression algorithm
- Lastly, we select the method by which the model performance can be evaluated.



Running and Evaluating the Model

accuracy: 84.69%

	true 0	true 1	class precision
pred. 0	917	162	84.99%
pred. 1	6	12	66.67%
class recall	99.35%	6.90%	


However, even though the overall accuracy is excellent, the accuracy for actual positive diagnosis is relatively poor, as can be seen in the above snapshot. In order to deal with this issue, we will recalibrate and enhance the model performance by implementing optimizing techniques.

Optimization and Recalibration of the Model

Optimization is one of the important steps before the deployment of any model, hence we have made use of threshold technique. In order to implement the threshold, we created and applied on the attribute containing the soft data, which in our case is 10YRSCHD.

Threshold Technique & Result

Parameters

 Create Threshold

threshold

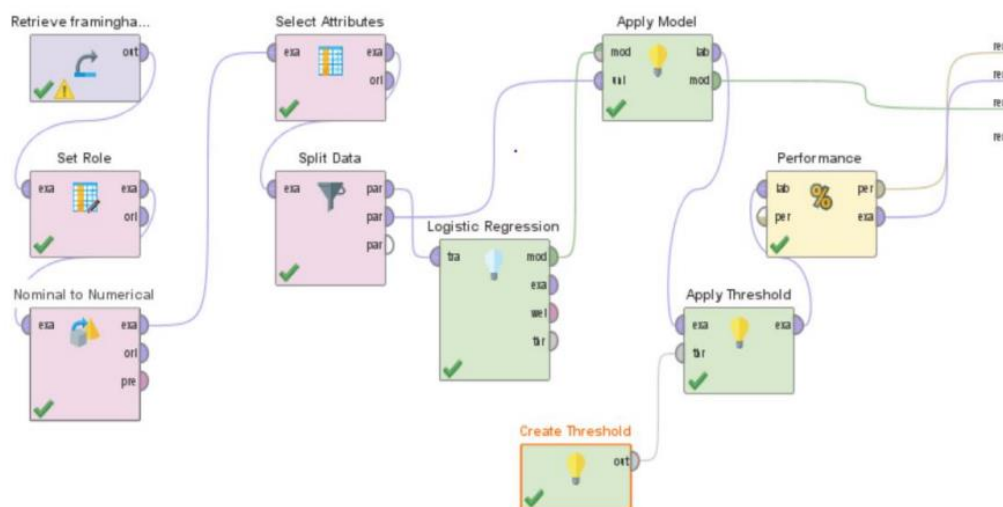
0.15

first class

0

second class

1



Results of the optimization by implementing threshold with different values

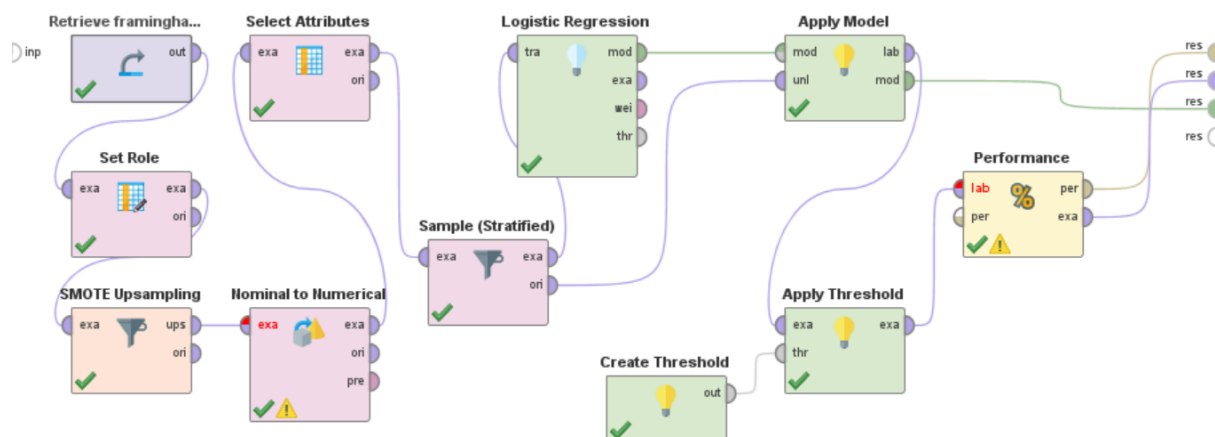
For Threshold Value = 0.5 accuracy: 84.69%				For Threshold Value = 0.4 accuracy: 83.32%			
	true 0	true 1	class precision		true 0	true 1	class precision
pred. 0	917	162	84.99%	pred. 0	895	155	85.24%
pred. 1	6	12	66.67%	pred. 1	28	19	40.43%
class recall	99.30%	5.90%		class recall	96.97%	10.92%	

For Threshold Value = 0.3 accuracy: 86.46%				For Threshold Value = 0.2 accuracy: 75.11%			
	true 0	true 1	class precision		true 0	true 1	class precision
pred. 0	840	132	86.42%	pred. 0	737	87	89.44%
pred. 1	83	42	33.60%	pred. 1	166	97	31.87%
class recall	91.01%	24.14%		class recall	79.85%	50.00%	

For Threshold Value = 0.15 accuracy: 67.73%				For Threshold Value = 0.1 accuracy: 53.87%			
	true 0	true 1	class precision		true 0	true 1	class precision
pred. 0	625	56	91.78%	pred. 0	446	31	93.53%
pred. 1	299	118	28.37%	pred. 1	475	143	23.14%
class recall	67.71%	67.82%		class recall	48.54%	62.18%	

Smote Up sampling and its Outcome

SMOTE is an oversampling approach that creates synthetic minority class samples. It is also used to create a synthetically or virtually class-balanced training set that is then utilized to build a classifier.



Result of the smote up sampling can be seen in below table,

Without SMOTE:

accuracy: 81.69%			
	true 0	true 1	class precision
pred 0	295	61	82.87%
pred 1	6	4	40.00%
class recall	98.01%	6.15%	

With SMOTE:

accuracy: 68.71%			
	true 0	true 1	class precision
pred 0	207	85	70.89%
pred 1	109	219	66.77%
class recall	65.51%	72.04%	

Documentation of Research results

If we compare the above two results, the accuracy obtained from smote up sampling is much better than that of the threshold technique. But, the advantage of smote up sampling is that it always results into creation of values that are artificial. This makes it doubtful to accept and work with live scenario use cases like detecting CHD after a decade. Thus, we decided to use the threshold technique over smote up sampling.

Assumptions on the result

Below are the result assumptions based on trained model

- Ten years coronary heart disease (10YEARSCHD) is not affected by the continuous values.
- Among all ratios, the 70:30 data split is the best.

Limitations of the approach

- The limitation of using classification algorithm is that it can only predict whether or not something will happen, but not the likelihood or percentage of that event occurring.

Constraints

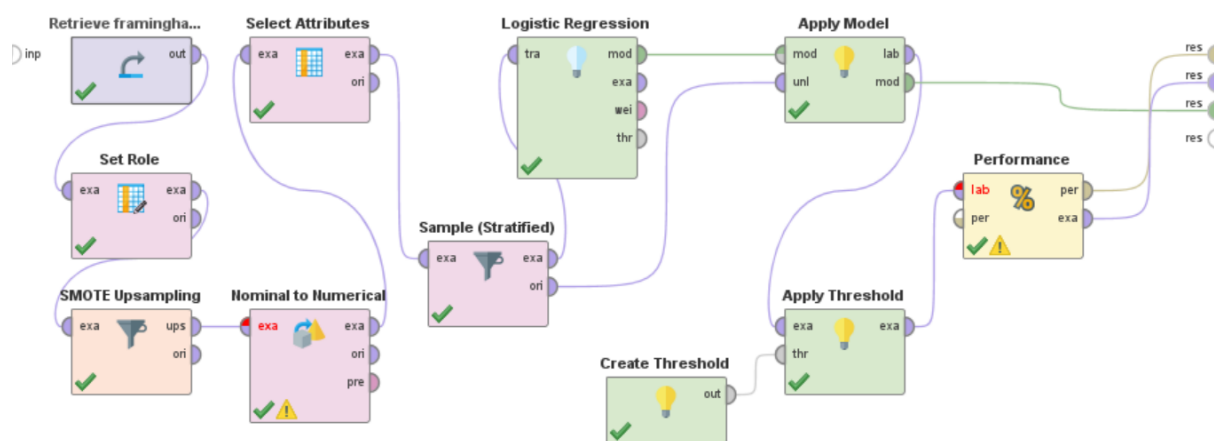
- The overall model's accuracy is quite standard (~ 68 percent).

Deployment

The deployment of the model is the final stage, now the deployment can depend on several ranging areas like it can be used in maintenance activity, can be deployed on cloud making it available to others for the usage.

Model Business Validation

To perform the business validation, we have made use of STRATIFIED SAMPLING. Stratified random sampling is a sampling method in which a population is divided into smaller common section called strata. Strata are produced in stratified random sampling, or stratification, depending on unique qualities among individuals, such as revenue or academic achievement.



Conclusion and Recommendations

We may conclude that by adopting a similar strategy, technology, and tools like Rapid Miner and Python to create Prediction Models for other cities and serve the nation and humanity, we can implement Prediction Models for other similar data as well. In this study, we examine the performance of multiple machine learning regression models in order to determine the optimal model for detecting the CHD possibilities in next ten years for a particular set of individuals.

accuracy: 68.00%			
	true 0	true 1	class precision
pred. 0	209	18	92.07%
pred. 1	94	29	23.58%
class recall	68.98%	61.70%	

- We need some more data to predict the accuracy reliably.
- Cross-validation with Python would also be beneficial. Other cross validation procedures should be used.
- Its application could be broadened to other industries as well.

THANKYOU!! 