Heart Disease (Predicted and Observed Cardiovascular Diseases): Implementing MachineLearning Algorithms on Framingham Dataset

Doris Allamani[a,*], Ekaterina Dubkova[b,*], Jean Rene Kenmegne Tchuente[c,*]

[a]*University Paris 1 Pantheon Sorbonne, France*
[b]*University Paris 1 Pantheon Sorbonne, France*
[c]*University Paris 1 Pantheon Sorbonne, France*

1. Introduction

Our Research objective aims at predicting the coronary heart disease in the next 10 yearsin other to determine if the observant will have it or not.

Within the framework of this study, we use the following machine learning algorithms: Logistic regression, Naïve Bayes, Random Forest and X-boosting. Furthermore, we use the oversampling methods to make the dataset more balanced which significantly increases the recall of the models used. These approaches enable us to explore the Framingham dataset on a wider scope and determine which once get to be more accurate predicting the heart disease.

Our motivation leading to us doing this project was centered around the fact that results from this analysis will enable policymakers in the financial sector to best know their clients' status as far as a loan or insurance contract could be proposed to them.

Framingham Risk Score

The Framingham dataset was collected in Massachusetts, US, and its Risk score is one of the first predictive scores for CHD. This was based on the Framingham Heart study examinations of 1971 to 1974, which included participants from either the original Framinghamstudy or from the initial investigation of the Framing- ham Offspring study.

Included subjects (n = 5345) were between 30 to 74 years old and free of CVD. All participants were followed-up for 12 years to ascertain the occurrence of CHD (angina pectoris, recognized and unrecognized MI, coronary insufficiency and CHD death). Hard CHD events included CHD death and MI.

In 1998, Wilson et al. developed a sex-specific prediction algorithm to estimate 10-year CHD risk by relating the Fifth Report of the Joint National Committee on Detection, Evaluation and Treatment of High Blood Pressure blood pressure and National Cholesterol Education Program cholesterol categories with age, the presence of T2DM and smoking.

In 2008, D'Agostino et al., based on a larger cohort of Framingham study, formulateda new sex-specific risk function tool that assessed not only the 10-year probability of CHD events, but also the risk for a first cardiovascular event (CHD, stroke, intermittent claudication and congestive heart failure).

[*]Corresponding author
Email addresses: d.allamani99@gmail.com (Doris Allamani), ekaterinadubkova06@gmail.com (Ekaterina Dubkova), Jean-Rene.Kenmegne-Tchuente@etu.univ-paris1.fr (Jean Rene Kenmegne Tchuente)

January 8, 2023

Although many concerns have been raised regarding the applicability and validity of this risk tool in different and diverse populations, many studies have validated it in other populations. Other versions of Framingham risk score have also been developed, including the Lifetime Framingham CVD Risk Score at 50 years of age and the 30-year Framingham cardiovascular risk score.

Even though the Framingham risk score is of the first predictive scores for CHD, it has been updated by the introduction of the Pooled Cohort Equations, which incorporates the Framingham study cohort. Therefore, the use of the Framingham risk score is not currently recommended.

2. Data Description & Methodology

Cardiovascular disease (CVD) is the leading cause of death worldwide. Management of cardiovascular risk factors, particularly hypertension and dyslipidemia has been shown to reduce cardiovascular morbidity and mortality.

However, current guidelines recommend adjusting the intensity of blood pressure and lipid-lowering treatment according to the cardiovascular risk of the patient.

Therefore, cardiovascular risk prediction is mandatory for optimizing cardiovascular prevention strategies, particularly in patients without established CVD or type 2 diabetes mellitus (T2DM). As a result, several cardiovascular risk prediction equations have been developed. Nevertheless, it is still unclear which is the optimal prediction risk equation.

In this section we describe the type of data we will be working with and their different categories, such as the Target Variable, Medical Data (Historical & Current), as well as Demographic and Behavioral Characteristics. We count 3 Demographic variables (Sex, Age, Education) of our sample, 2 Behavioral variables, 4 Medical (historical) indicators, 6 Medical (current) indicators with CHD as a target variable giving us a total of 16 indicators in the data set with a sample of 4238 observations.

Demographic

- Sex: male or female, 1 and 0 respectively. (Nominal)
- Age: age of the patient. (Continuous)
- Education: four levels of education: 1 - less than high school, 2 - high school graduates, 3 - college graduates, 4 - post-college graduates. (Ordinal)

Behavioral

- Current Smoker: whether or not the patient is a current smoker, 1 and 0 respectively. (Nominal)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (Continuous)

Medical(history)

- BP Meds: whether or not the patient was on blood pressure medication, 1 and 0 respectively. (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke, 1 and 0 respectively. (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive, 1 and 0. (Nominal)
- Diabetes: whether or not the patient had diabetes, 1 and 0 respectively. (Nominal)

Medical(current)

- Tot Chol: total cholesterol level. (Continuous)
- Sys BP: systolic blood pressure. (Continuous)
- Dia BP: diastolic blood pressure. (Continuous)
- BMI: Body Mass Index. (Continuous)
- Heart Rate: heart rate. (Continuous)
- Glucose: glucose level. (Continuous)

Target variable

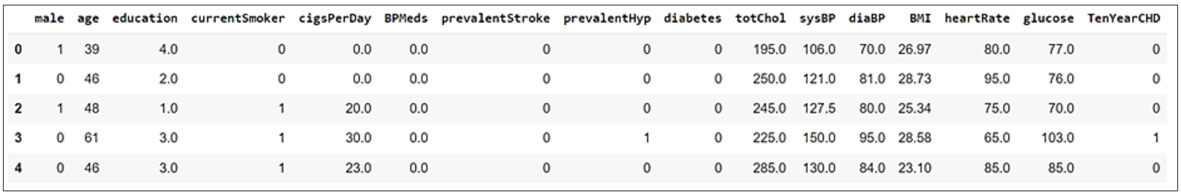
- Ten Year CHD: 10 year risk of coronary heart disease CHD, 1 - if yes, 0 - if no. (Nominal)

3. Data Analysis

Descriptive Statistics

In this section we aim at first exploring the central limit theory of our data set, that is looking at the mode, mean and standard deviation. Alongside, identifying the missing values which we needed to fill in order to be able to normalize our dataset.

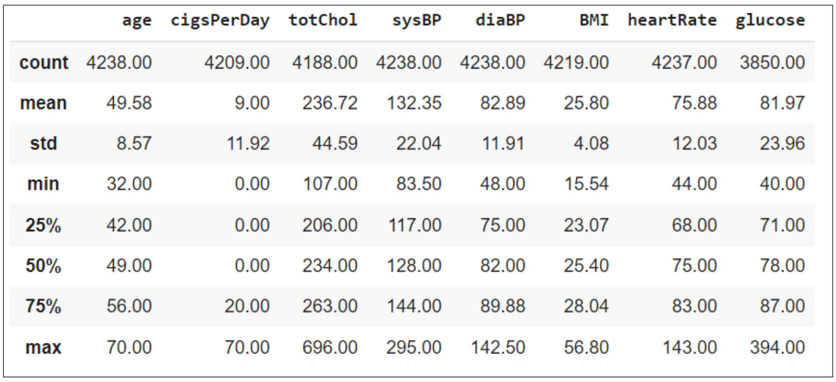
As mentioned previously, we are working on 16 indicators in the data set with a sample of 4238 observations as shown below:



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Table 1: Data overview

The main descriptive statistics for continuous variables is presented below:



	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
count	4238.00	4209.00	4188.00	4238.00	4238.00	4219.00	4237.00	3850.00
mean	49.58	9.00	236.72	132.35	82.89	25.80	75.88	81.97
std	8.57	11.92	44.59	22.04	11.91	4.08	12.03	23.96
min	32.00	0.00	107.00	83.50	48.00	15.54	44.00	40.00
25%	42.00	0.00	206.00	117.00	75.00	23.07	68.00	71.00
50%	49.00	0.00	234.00	128.00	82.00	25.40	75.00	78.00
75%	56.00	20.00	263.00	144.00	89.88	28.04	83.00	87.00
max	70.00	70.00	696.00	295.00	142.50	56.80	143.00	394.00

Table 2: Description of Continuous variables

As we can see from the *table 2*, the sample consists of the observants from 32 to 70 years old, the average age is 50 years with almost 9 years standard deviation. The average blood pressure is a little bit higher than the normal (120/70 or 130/80), the total cholesterol 236.72 mg/dL close to the limit, whereas heart rate is between the normal values of 60 and 80 beats per minute and glucose is 82 mg/dL, under the maximum norm.

The pie charts, *figure 1*, display the categorical indicators where we observed mainly an even distribution of both males and females, education characteristics and current smokers (whether the person is current smoker or not).

However, we can notice that BP Meds, Prevalent stroke, and Diabetes have a very low proportion with the amount of people getting the illness. This could be partly explain to the small sample of observations compared to the total US population. Another reason for the unbalanced observations in terms of the characteristics above is that the actual statistics shows that less than 10% of population have heart diseases and diabetes.

Therefore, the sample representing real world data could not be balanced a priory, which we can also see from the pie charts below:

Pie Charts for characteristic variables

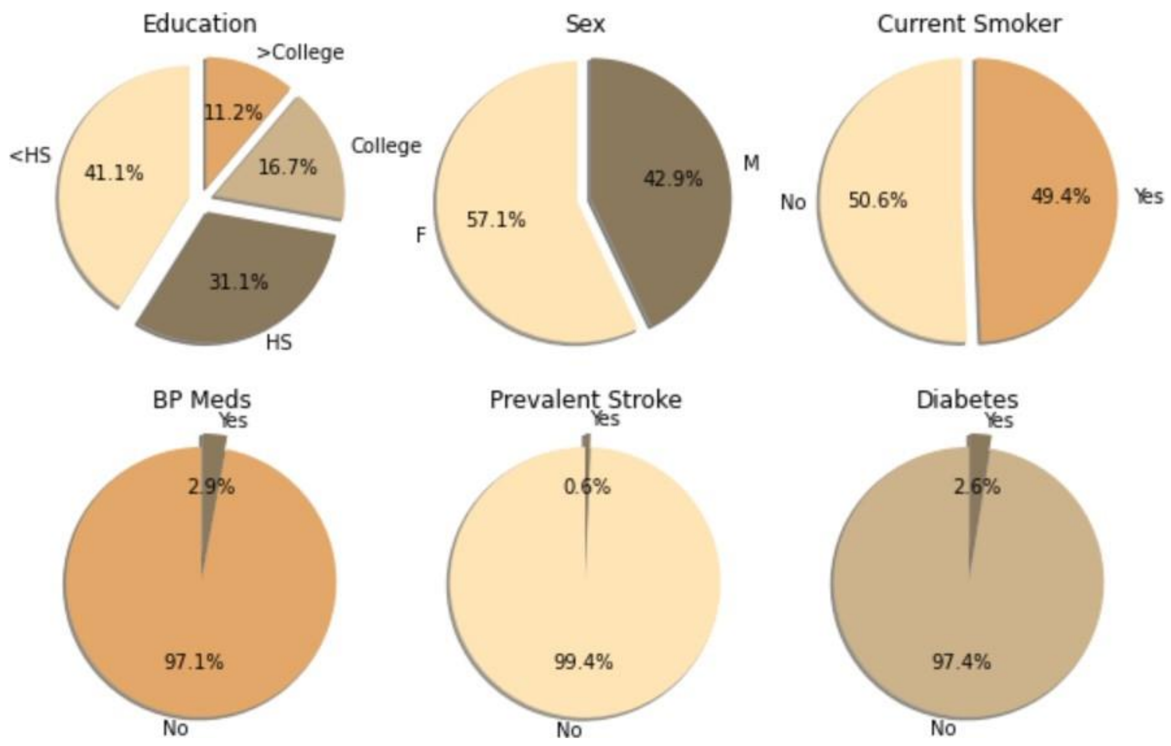


Figure 1: Pie charts for categorical data

We note most of the observants are females, non-smokers, who didn't have strokes before, with average heart rate and cholesterol levels, who don't have the risk of coronary heart disease happening in the next 10 years (only 644 observants had a 10-year heart disease, whereas 3594 of them didn't).

3.1. Missing Data

The next step of our analysis is dealing with the missing values. There are 105 points missing from education variable (ordinal variable taking values 1 to 4), 29 points from cigsPerDay, 53 points from BPMeds, 50 from totChol, 19 from BMI and 388 from glucose. We can note that 9% of the glucose data is missing. To fill the missing values, we use the kNN (k nearest neighbors) algorithm with the number of neighbors equal to 5, but before running the algorithm, it is very important to normalize the data (continuous variables). The technique used for the normalization process is MinMax, with the below formula:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

We also rounded the outcomes on the categorical data which was filled with non-integer values after implementing the kNN.

After normalizing the data and filling the missing values with kNN method, we can see that the continuous variables are normally distributed with some positive skewness (long right tales, especially for total cholesterol and glucose characteristics), *Figure 2*.

Histograms on current medical data

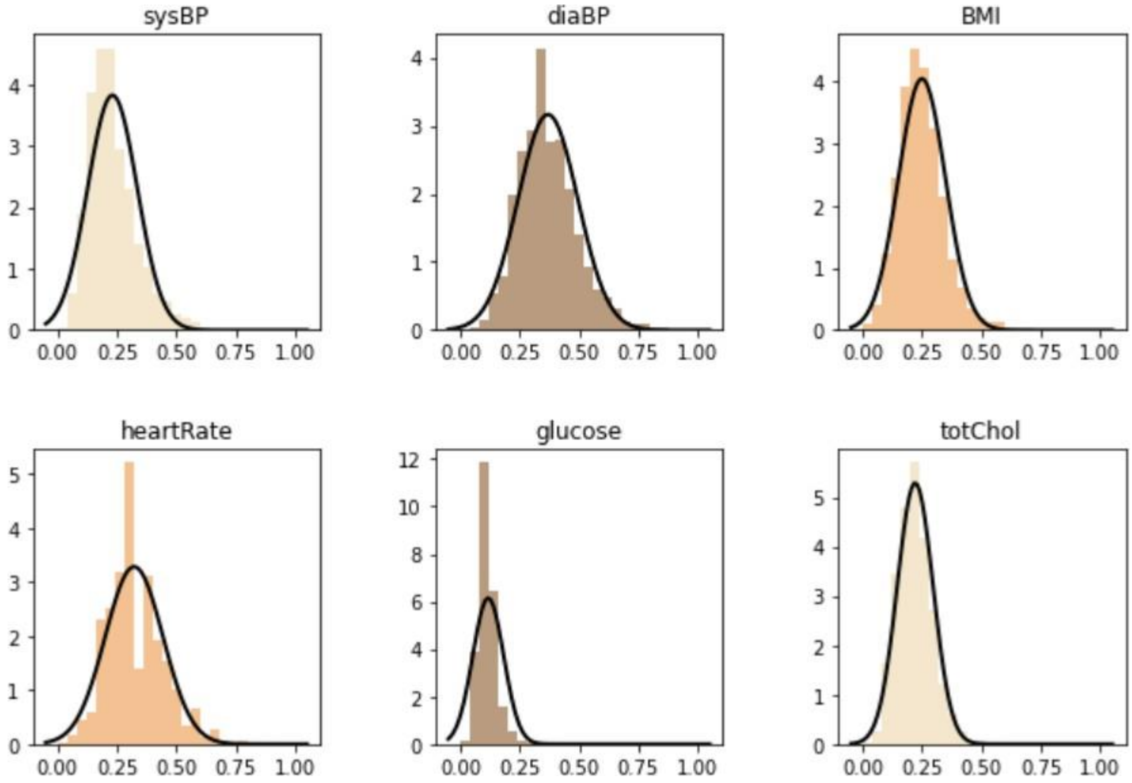
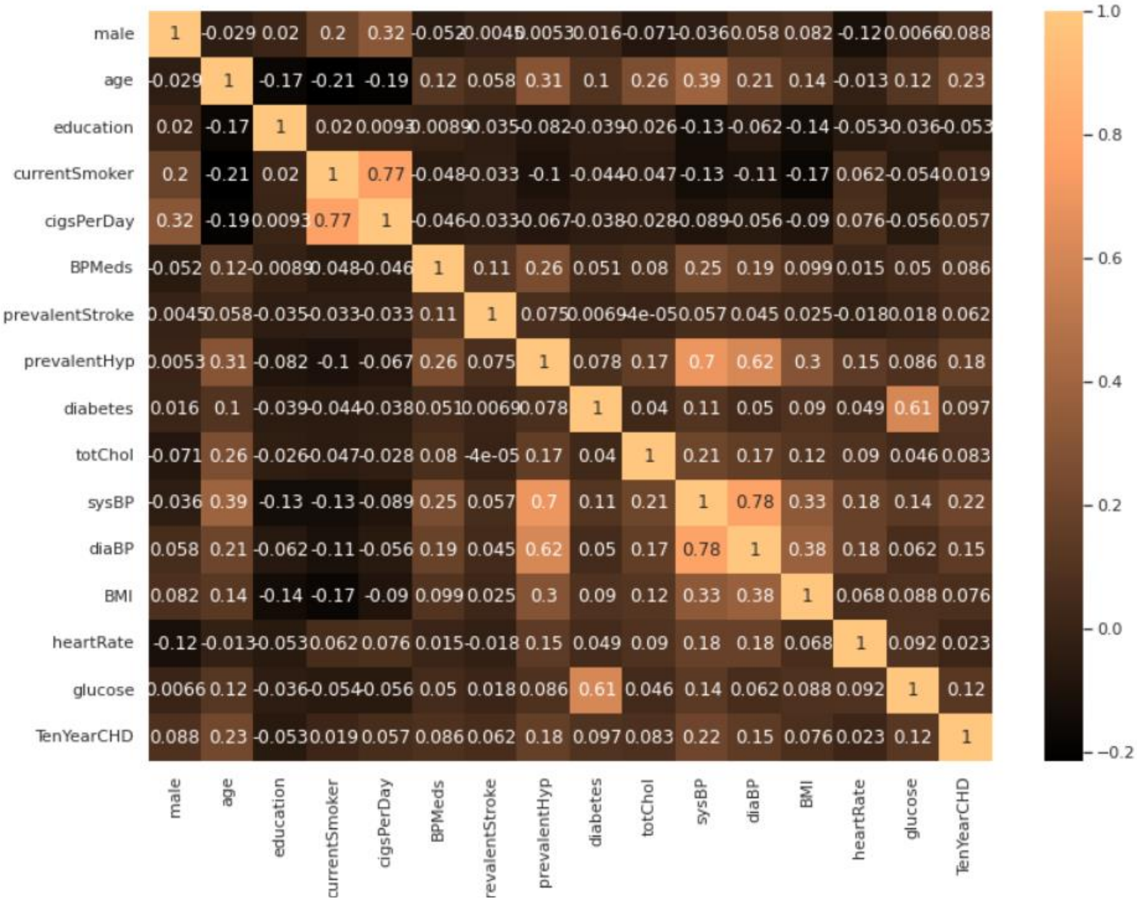


Figure 2: Current Medical Data Distributions

3.2. Variable Selection

We follow the analysis with a heat map, *table 3*, which permitted us to understand the correlation analysis within the variables. Here, we pay attention to very light (positive) or very dark (negative) cells since they represent high correlation level within variables such as the indicator cigarettes per day with the current smoker (0.77) or sysBP with diaBP (0.78) as shown below:

Heat Map



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
male	1	-0.029	0.02	0.2	0.32	-0.052	0.0045	0.0053	0.016	-0.071	-0.036	0.058	0.082	-0.12	0.0066	0.088
age	-0.029	1	-0.17	-0.21	-0.19	0.12	0.058	0.31	0.1	0.26	0.39	0.21	0.14	-0.013	0.12	0.23
education	0.02	-0.17	1	0.02	0.0093	-0.0089	-0.035	-0.082	-0.039	-0.026	-0.13	-0.062	-0.14	-0.053	-0.036	-0.053
currentSmoker	0.2	-0.21	0.02	1	0.77	-0.048	-0.033	-0.1	-0.044	-0.047	-0.13	-0.11	-0.17	0.062	-0.054	0.019
cigsPerDay	0.32	-0.19	0.0093	0.77	1	-0.046	-0.033	-0.067	-0.038	-0.028	-0.089	-0.056	-0.09	0.076	-0.056	0.057
BPMeds	-0.052	0.12	-0.0089	-0.048	-0.046	1	0.11	0.26	0.051	0.08	0.25	0.19	0.099	0.015	0.05	0.086
prevalentStroke	0.0045	0.058	-0.035	-0.033	-0.033	0.11	1	0.075	0.0069	-4e-05	0.057	0.045	0.025	-0.018	0.018	0.062
prevalentHyp	0.0053	0.31	-0.082	-0.1	-0.067	0.26	0.075	1	0.078	0.17	0.7	0.62	0.3	0.15	0.086	0.18
diabetes	0.016	0.1	-0.039	-0.044	-0.038	0.051	0.0069	0.078	1	0.04	0.11	0.05	0.09	0.049	0.61	0.097
totChol	-0.071	0.26	-0.026	-0.047	-0.028	0.08	-4e-05	0.17	0.04	1	0.21	0.17	0.12	0.09	0.046	0.083
sysBP	-0.036	0.39	-0.13	-0.13	-0.089	0.25	0.057	0.7	0.11	0.21	1	0.78	0.33	0.18	0.14	0.22
diaBP	0.058	0.21	-0.062	-0.11	-0.056	0.19	0.045	0.62	0.05	0.17	0.78	1	0.38	0.18	0.062	0.15
BMI	0.082	0.14	-0.14	-0.17	-0.09	0.099	0.025	0.3	0.09	0.12	0.33	0.38	1	0.068	0.088	0.076
heartRate	-0.12	-0.013	-0.053	0.062	0.076	0.015	-0.018	0.15	0.049	0.09	0.18	0.18	0.068	1	0.092	0.023
glucose	0.0066	0.12	-0.036	-0.054	-0.056	0.05	0.018	0.086	0.61	0.046	0.14	0.062	0.088	0.092	1	0.12
TenYearCHD	0.088	0.23	-0.053	0.019	0.057	0.086	0.062	0.18	0.097	0.083	0.22	0.15	0.076	0.023	0.12	1

Table 3: Correlation Matrix (Pearson)

In some of our models we exclude these variables.

3.3. Outliers

After normalizing the data in previous step, and standardizing it, we now look for outliers which are defined as observations that deviate from the main portion of observations by a significant magnitude.

Below we have plotted the Principal Component Analysis (PCA) on two dimesions, *figure 3*. This algorithm help us to demonstate our multivariate data into a more comprehensive form. First of all, we observe that there are no clusters formed into our dataset. Secondly, we notice that some points are in a further distance than the rest.

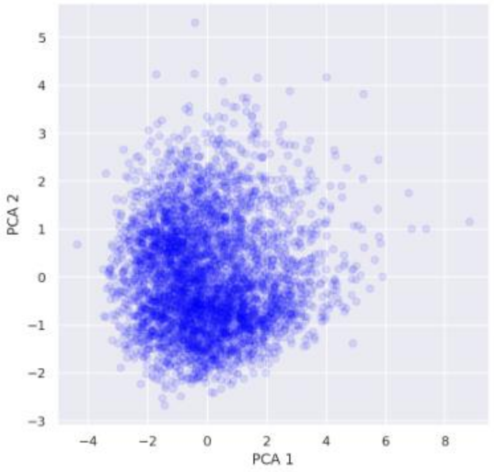


Figure 3: PCA reduced in two dimensions

The boxplot of the continuos variables, *figure 4,* caution us that indeed there are a lot outliers. We have decided to not omitt any of them as the dataset is unbalanced regarding the target variable, the missing data were filled with kNN and these observants may be rare medical cases with abnormal medical parameters.

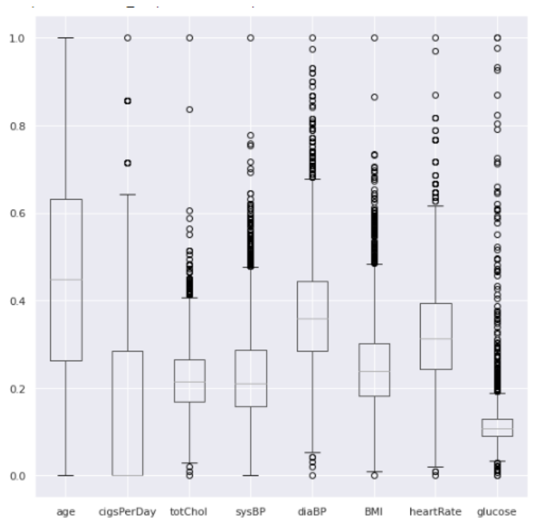
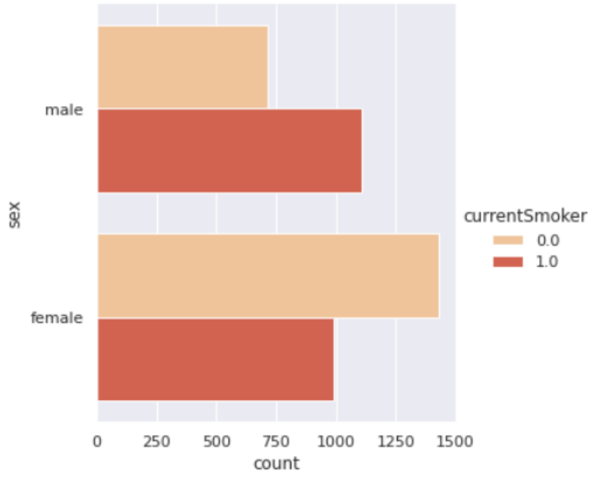


Figure 4: Box plot of the continuous variables

3.4. Further analysis

Some more descriptive statistics:



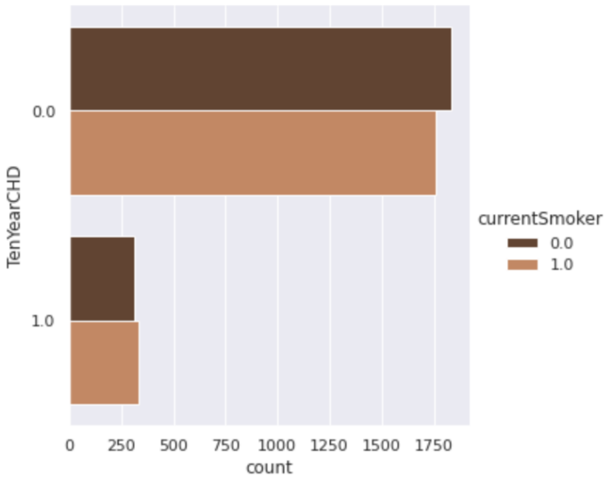


Figure 5: Smokers by sex and effect on TenYearCHD

Given the graphs above, we observe that there are more smokers who have the heart disease than non-smokers, nevertheless, there amounts are almost equal. Higher share of males than females are current smokers.

4. Logistic Regression

4.1. Logistic regression without oversampling

Given our research objective and the type of data we are working on, one of the classifier models we have choose to implement is the Logistic regression. This type of statistical model (also known as logit model) is often used for classification and predictive analytics. With this regression, we aim to estimate the probability of heat diseases occurring, based on our given dataset of independent variables. Since the outcome is a probability, our Target variable (10 Years heart disease) variable is bounded between 0 and 1. Our logistic transformation is applied on the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formula:

$$P(y=1) = \frac{1}{1+exp^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

The metrics we have chosen to validate our models are:

Accuracy represents the number of correctly classified data instances over the total number of data instances

Metrics:

- Accuracy - measures the number of correctly classifies data instances over total number of data instances

- Precision - the share of instances called "positive" by the algorithm and actually beingpositive.

- Recall - the share of positive instances which algorithm has actually found among all the positive instances.

- F1 – takes into the account precision and recall and reaches a maximum with precision and recall equal to one and isclose to zero if one of the arguments is close to zero.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

To build the Logistic Regression model, we explore different variants of grouping the features. We have grouped the variables as below:

1. Only demographical and medical(history) parameters

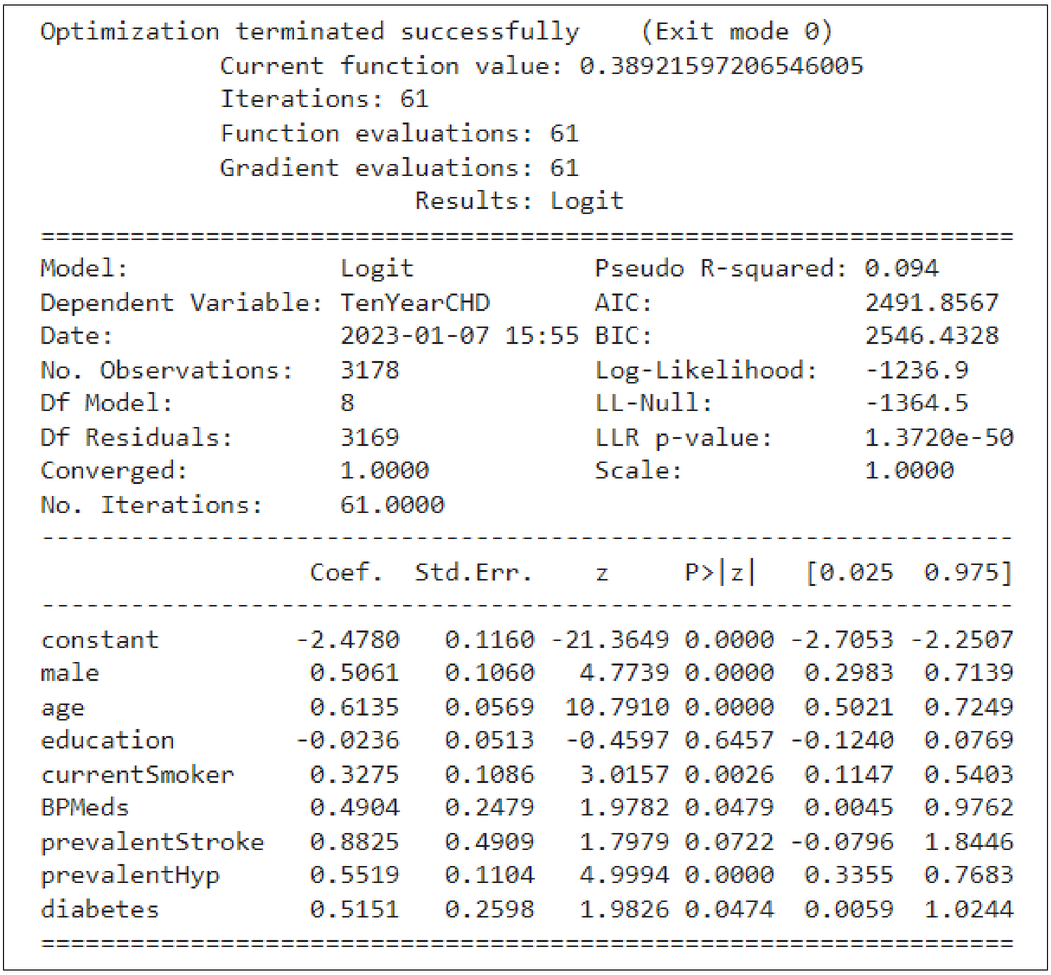
2. Only demographical and medical(current) parameters

3. Only uncorrelated parameters (excluding sysBP and cigsPerDay)

4. All the parameters in the dataset

Logistic Regression 1

In the first logistic regression, we estimated the effect on TenYearCHD of only demographical and medical (historical) parameters as shown in our data description referenced above. To proceed with the logistic regression, we divide the dataset into train and test parts (test part taking 25% of all the points) and after running this subset of our data in our parameterized logistic regression formula we obtained the following results:



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.38921597206546005
            Iterations: 61
            Function evaluations: 61
            Gradient evaluations: 61
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.094
Dependent Variable: TenYearCHD       AIC:              2491.8567
Date:               2023-01-07 15:55 BIC:              2546.4328
No. Observations:   3178             Log-Likelihood:   -1236.9
Df Model:           8                LL-Null:          -1364.5
Df Residuals:       3169             LLR p-value:      1.3720e-50
Converged:          1.0000           Scale:            1.0000
No. Iterations:     61.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
constant         -2.4780   0.1160 -21.3649 0.0000 -2.7053 -2.2507
male              0.5061   0.1060   4.7739 0.0000  0.2983  0.7139
age               0.6135   0.0569  10.7910 0.0000  0.5021  0.7249
education        -0.0236   0.0513  -0.4597 0.6457 -0.1240  0.0769
currentSmoker     0.3275   0.1086   3.0157 0.0026  0.1147  0.5403
BPMeds            0.4904   0.2479   1.9782 0.0479  0.0045  0.9762
prevalentStroke   0.8825   0.4909   1.7979 0.0722 -0.0796  1.8446
prevalentHyp      0.5519   0.1104   4.9994 0.0000  0.3355  0.7683
diabetes          0.5151   0.2598   1.9826 0.0474  0.0059  1.0244
=================================================================
```

Figure 6: Logistic Regression 1

As we can see, all the regressors (except of education) are significant on 10% significance level. Nevertheless, omitting the education variable from the list of regressors doesn't change the metrics of the model:

The confusion matrix is:

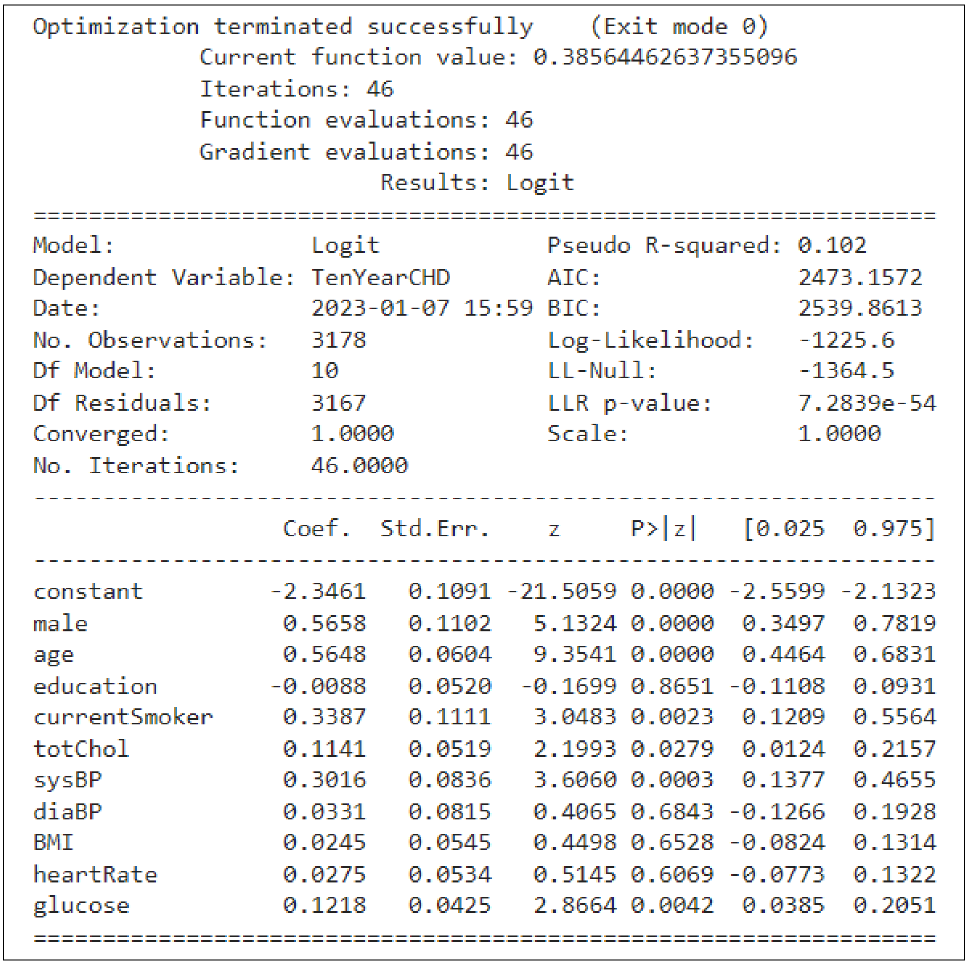
[[903   2]
[146   9]]

Accuracy: 0.86
Precision: 0.82
Recall: 0.06
F1: 0.11

The model achieves pretty high accuracy and precision (both above 80%), but at the same time the value of recall is very low (6%), which means that the algorithm finds positive observants among all the positive observant not good enough (and for us it's important to predict specifically positive observants correctly).

Logistic Regression 2

In the Second logistic regression, we grouped the demographical and medical (current) parameters. We obtained the following results:



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.38564462637355096
            Iterations: 46
            Function evaluations: 46
            Gradient evaluations: 46
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.102
Dependent Variable: TenYearCHD       AIC:              2473.1572
Date:               2023-01-07 15:59 BIC:              2539.8613
No. Observations:   3178             Log-Likelihood:   -1225.6
Df Model:           10               LL-Null:          -1364.5
Df Residuals:       3167             LLR p-value:      7.2839e-54
Converged:          1.0000           Scale:            1.0000
No. Iterations:     46.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
constant         -2.3461   0.1091 -21.5059 0.0000 -2.5599 -2.1323
male              0.5658   0.1102   5.1324 0.0000  0.3497  0.7819
age               0.5648   0.0604   9.3541 0.0000  0.4464  0.6831
education        -0.0088   0.0520  -0.1699 0.8651 -0.1108  0.0931
currentSmoker     0.3387   0.1111   3.0483 0.0023  0.1209  0.5564
totChol           0.1141   0.0519   2.1993 0.0279  0.0124  0.2157
sysBP             0.3016   0.0836   3.6060 0.0003  0.1377  0.4655
diaBP             0.0331   0.0815   0.4065 0.6843 -0.1266  0.1928
BMI               0.0245   0.0545   0.4498 0.6528 -0.0824  0.1314
heartRate         0.0275   0.0534   0.5145 0.6069 -0.0773  0.1322
glucose           0.1218   0.0425   2.8664 0.0042  0.0385  0.2051
=================================================================
```

Figure 7: Logistic Regression 2

The confusion matrix is:

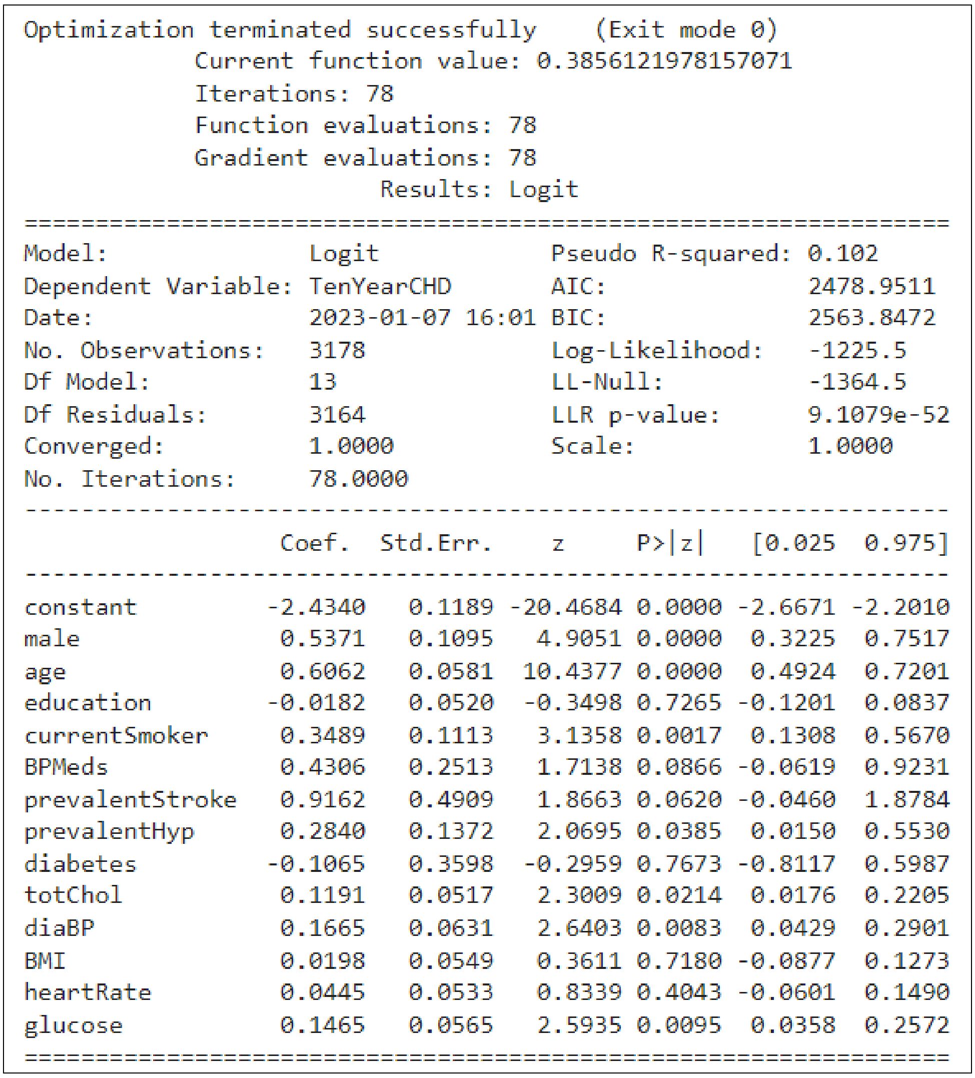
```
[[903   2]
 [147   8]]
```

Accuracy: 0.86
Precision: 0.80
Recall: 0.05
F1: 0.10

The results obtained are very similar to the first logistic regression. Deleting insignificant variables, such as education, diaBP and BMI doesn’t improve the metrics of the model.

Logistic Regression 3

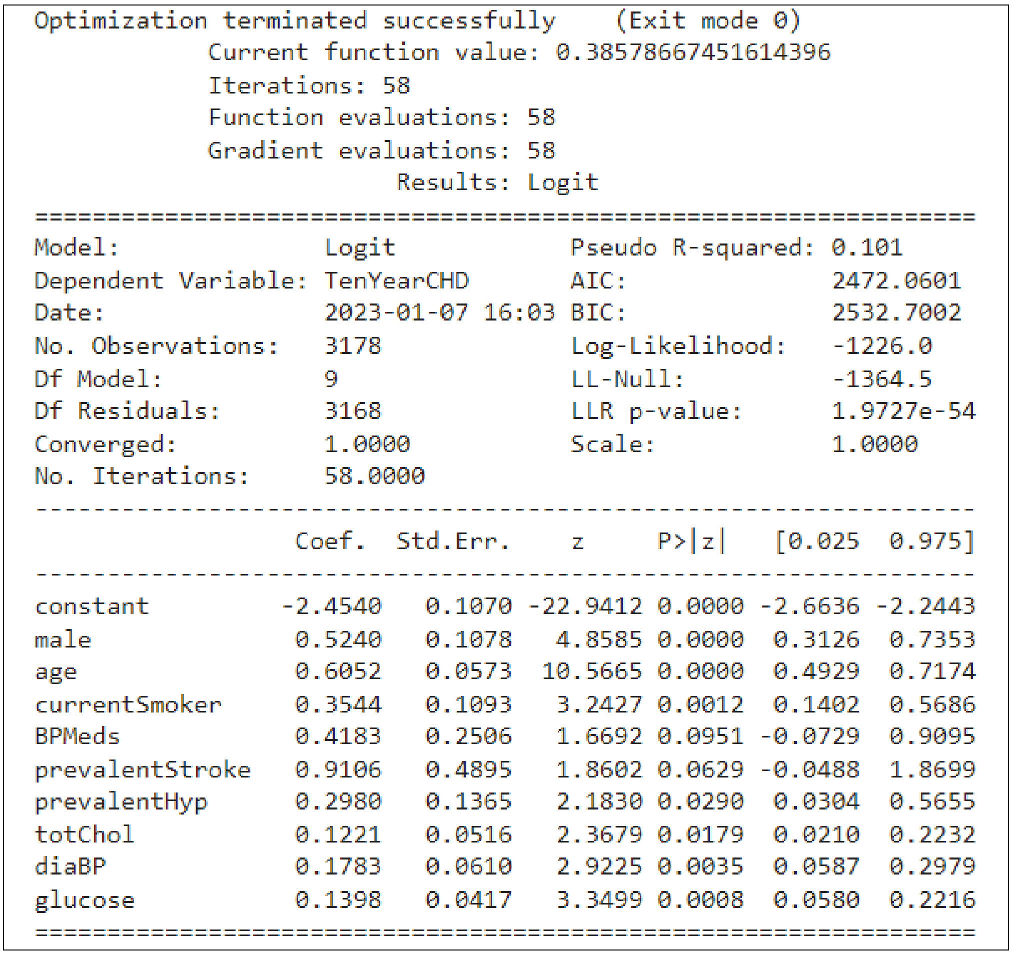
In the Third logistic regression, we grouped the uncorrelated parameters (see the heat map, *table 3*) parameters as shown in our data description referenced above and after running this subset of our data in our parameterized logistic regression formula, we obtained the following results:



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.3856121978157071
            Iterations: 78
            Function evaluations: 78
            Gradient evaluations: 78
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.102
Dependent Variable: TenYearCHD       AIC:              2478.9511
Date:               2023-01-07 16:01 BIC:              2563.8472
No. Observations:   3178             Log-Likelihood:   -1225.5
Df Model:           13               LL-Null:          -1364.5
Df Residuals:       3164             LLR p-value:      9.1079e-52
Converged:          1.0000           Scale:            1.0000
No. Iterations:     78.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
constant         -2.4340   0.1189 -20.4684 0.0000 -2.6671 -2.2010
male              0.5371   0.1095   4.9051 0.0000  0.3225  0.7517
age               0.6062   0.0581  10.4377 0.0000  0.4924  0.7201
education        -0.0182   0.0520  -0.3498 0.7265 -0.1201  0.0837
currentSmoker     0.3489   0.1113   3.1358 0.0017  0.1308  0.5670
BPMeds            0.4306   0.2513   1.7138 0.0866 -0.0619  0.9231
prevalentStroke   0.9162   0.4909   1.8663 0.0620 -0.0460  1.8784
prevalentHyp      0.2840   0.1372   2.0695 0.0385  0.0150  0.5530
diabetes         -0.1065   0.3598  -0.2959 0.7673 -0.8117  0.5987
totChol           0.1191   0.0517   2.3009 0.0214  0.0176  0.2205
diaBP             0.1665   0.0631   2.6403 0.0083  0.0429  0.2901
BMI               0.0198   0.0549   0.3611 0.7180 -0.0877  0.1273
heartRate         0.0445   0.0533   0.8339 0.4043 -0.0601  0.1490
glucose           0.1465   0.0565   2.5935 0.0095  0.0358  0.2572
=================================================================
```

Figure 8: Logistic Regression 3

And after deleting the insignificant variables (education, diabetes, BMI, heartrate):



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.38578667451614396
            Iterations: 58
            Function evaluations: 58
            Gradient evaluations: 58
                         Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.101
Dependent Variable: TenYearCHD       AIC:              2472.0601
Date:               2023-01-07 16:03 BIC:              2532.7002
No. Observations:   3178             Log-Likelihood:   -1226.0
Df Model:           9                LL-Null:          -1364.5
Df Residuals:       3168             LLR p-value:      1.9727e-54
Converged:          1.0000           Scale:            1.0000
No. Iterations:     58.0000
------------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
constant         -2.4540   0.1070 -22.9412 0.0000 -2.6636 -2.2443
male              0.5240   0.1078   4.8585 0.0000  0.3126  0.7353
age               0.6052   0.0573  10.5665 0.0000  0.4929  0.7174
currentSmoker     0.3544   0.1093   3.2427 0.0012  0.1402  0.5686
BPMeds            0.4183   0.2506   1.6692 0.0951 -0.0729  0.9095
prevalentStroke   0.9106   0.4895   1.8602 0.0629 -0.0488  1.8699
prevalentHyp      0.2980   0.1365   2.1830 0.0290  0.0304  0.5655
totChol           0.1221   0.0516   2.3679 0.0179  0.0210  0.2232
diaBP             0.1783   0.0610   2.9225 0.0035  0.0587  0.2979
glucose           0.1398   0.0417   3.3499 0.0008  0.0580  0.2216
==================================================================
```

Figure 9: Logistic Regression 3 (significant variables)

The confusion matrix is:

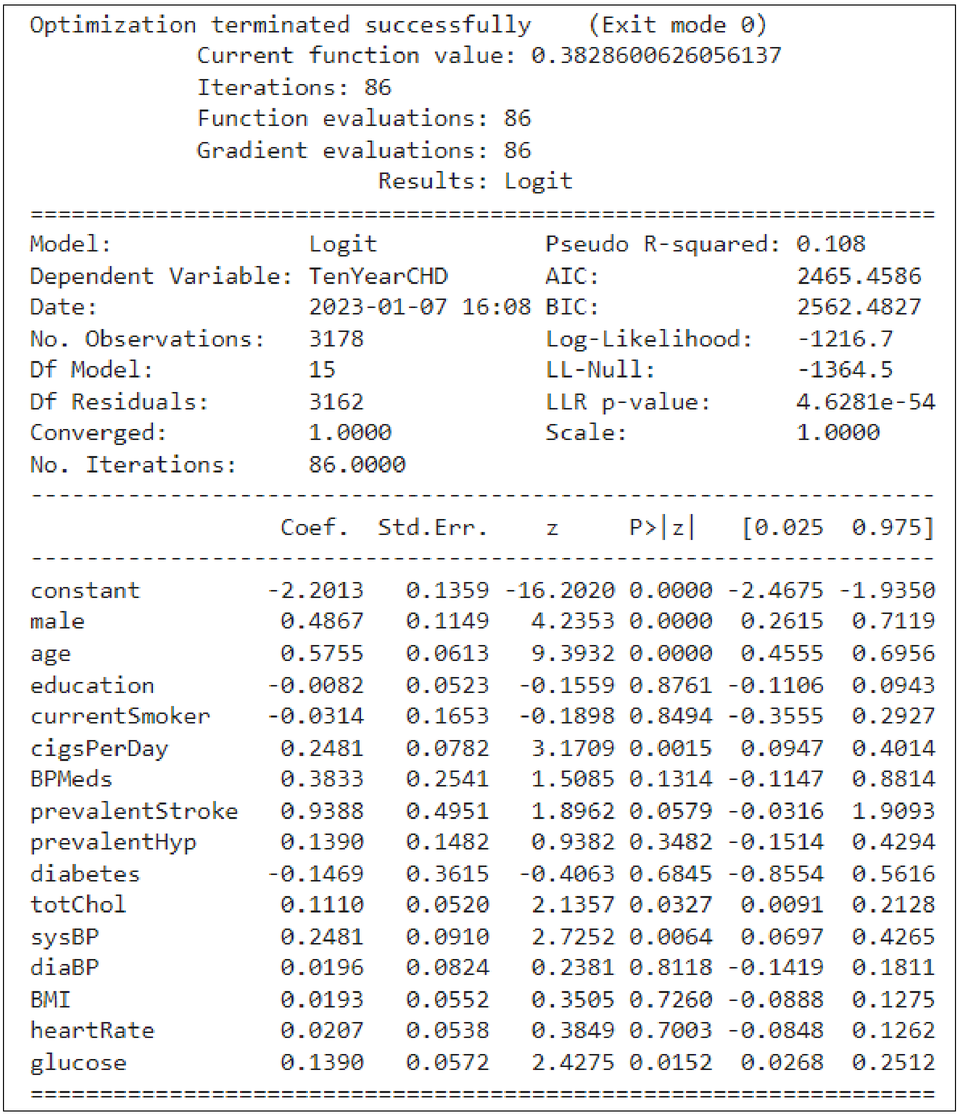
```
[[901   4]
 [147   8]]
```

Accuracy: 0.86
Precision: 0.67
Recall: 0.05
F1: 0.10

As we can notice, including more variables to the model (current medical together with historical medical) reduces precision, whereas accuracy and recall are staying on the same level. The problem still holds: the recall value is too low (5%), so we'll have to find alternative methods to increase this parameter.

Logistic Regression 4

In the fourth logistic regression, we have included all the variables despite their correlation. The summary of the results are as below:



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.3828600626056137
            Iterations: 86
            Function evaluations: 86
            Gradient evaluations: 86
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.108
Dependent Variable: TenYearCHD       AIC:              2465.4586
Date:               2023-01-07 16:08 BIC:              2562.4827
No. Observations:   3178             Log-Likelihood:   -1216.7
Df Model:           15               LL-Null:          -1364.5
Df Residuals:       3162             LLR p-value:      4.6281e-54
Converged:          1.0000           Scale:            1.0000
No. Iterations:     86.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
constant         -2.2013   0.1359 -16.2020 0.0000 -2.4675 -1.9350
male              0.4867   0.1149   4.2353 0.0000  0.2615  0.7119
age               0.5755   0.0613   9.3932 0.0000  0.4555  0.6956
education        -0.0082   0.0523  -0.1559 0.8761 -0.1106  0.0943
currentSmoker    -0.0314   0.1653  -0.1898 0.8494 -0.3555  0.2927
cigsPerDay        0.2481   0.0782   3.1709 0.0015  0.0947  0.4014
BPMeds            0.3833   0.2541   1.5085 0.1314 -0.1147  0.8814
prevalentStroke   0.9388   0.4951   1.8962 0.0579 -0.0316  1.9093
prevalentHyp      0.1390   0.1482   0.9382 0.3482 -0.1514  0.4294
diabetes         -0.1469   0.3615  -0.4063 0.6845 -0.8554  0.5616
totChol           0.1110   0.0520   2.1357 0.0327  0.0091  0.2128
sysBP             0.2481   0.0910   2.7252 0.0064  0.0697  0.4265
diaBP             0.0196   0.0824   0.2381 0.8118 -0.1419  0.1811
BMI               0.0193   0.0552   0.3505 0.7260 -0.0888  0.1275
heartRate         0.0207   0.0538   0.3849 0.7003 -0.0848  0.1262
glucose           0.1390   0.0572   2.4275 0.0152  0.0268  0.2512
=================================================================
```

Figure 10: Logistic Regression 4

The confusion matrix is:

```
[[902   3]
 [144  11]]

Accuracy: 0.86
Precision: 0.79
Recall: 0.07
F1: 0.13
```

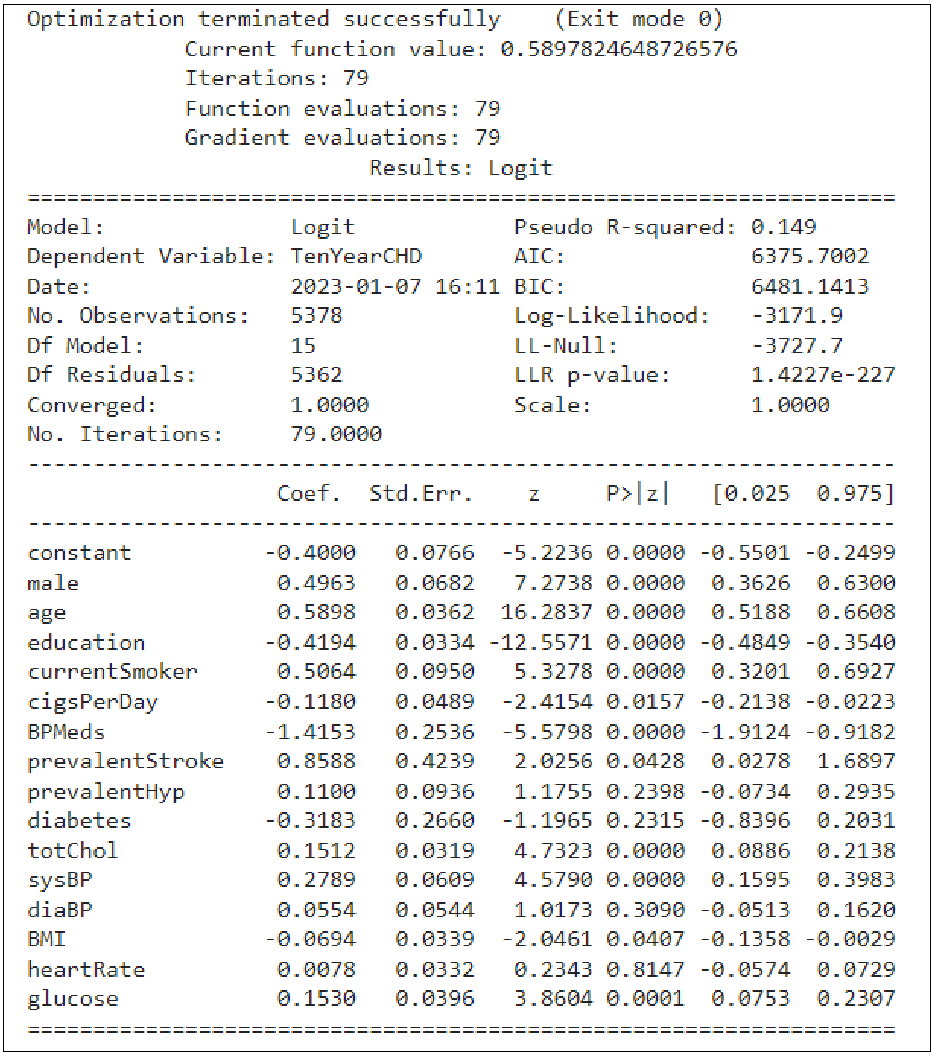
Using al the available data to predict the TenYearCHD turns out to be the best model in terms of those analyzed by logistic regression 1, 2 and 3. In all the four scenarios, despite the high accuracy and precision, we always observe an extremely low recall value (not more than 7%), which inspires us to look for the different ML algorithms to predict the target variable correctly.

4.2. Logistic regression with oversampling (using SMOTENC)

Proceeding further with the Logistic regression 4, we will use the SMOTENC oversampling algorithm in order to balance randomly our data train and achieve the results with higher recall value.

We kindly remind that the target variable is not balanced, therefore an oversampling technique could help our model to achieve better results. The SMOTENC algorithm stands short for Synthetic Minority Oversampling Technique for nominal and continuous, which works differently than SMOTE algorithm, as it accounts for nominal variables as well. It is important to note that this technique is only applied to the train dataset to avoid bias on the test set.

SMOTENC works by a k-nearest neighbors' algorithm to sample synthetic data. SMOTENC first starts by choosing random data from the minority class, then k-nearest neighbor's from the data are set. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. From this intuition, we redo the logistic regression and observe a more precise estimate:



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.5897824648726576
            Iterations: 79
            Function evaluations: 79
            Gradient evaluations: 79
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.149
Dependent Variable: TenYearCHD       AIC:              6375.7002
Date:               2023-01-07 16:11 BIC:              6481.1413
No. Observations:   5378             Log-Likelihood:   -3171.9
Df Model:           15               LL-Null:          -3727.7
Df Residuals:       5362             LLR p-value:      1.4227e-227
Converged:          1.0000           Scale:            1.0000
No. Iterations:     79.0000
------------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
constant          -0.4000   0.0766  -5.2236 0.0000 -0.5501 -0.2499
male               0.4963   0.0682   7.2738 0.0000  0.3626  0.6300
age                0.5898   0.0362  16.2837 0.0000  0.5188  0.6608
education         -0.4194   0.0334 -12.5571 0.0000 -0.4849 -0.3540
currentSmoker      0.5064   0.0950   5.3278 0.0000  0.3201  0.6927
cigsPerDay        -0.1180   0.0489  -2.4154 0.0157 -0.2138 -0.0223
BPMeds            -1.4153   0.2536  -5.5798 0.0000 -1.9124 -0.9182
prevalentStroke    0.8588   0.4239   2.0256 0.0428  0.0278  1.6897
prevalentHyp       0.1100   0.0936   1.1755 0.2398 -0.0734  0.2935
diabetes          -0.3183   0.2660  -1.1965 0.2315 -0.8396  0.2031
totChol            0.1512   0.0319   4.7323 0.0000  0.0886  0.2138
sysBP              0.2789   0.0609   4.5790 0.0000  0.1595  0.3983
diaBP              0.0554   0.0544   1.0173 0.3090 -0.0513  0.1620
BMI               -0.0694   0.0339  -2.0461 0.0407 -0.1358 -0.0029
heartRate          0.0078   0.0332   0.2343 0.8147 -0.0574  0.0729
glucose            0.1530   0.0396   3.8604 0.0001  0.0753  0.2307
==================================================================
```

Figure 11: Logistic Regression with oversampling

The confusion matrix is:

```
[[607 298]
 [ 61  94]]
Accuracy: 0.66
Precision: 0.24
Recall: 0.61
F1: 0.34
```

After implementing the SMOTENC algorithm, 94 observations in the test set were correctly predicted as positive, which results in recall of 61%. However, the accuracy and precision decreased by 20 p.p. and 40 p.p. respectively, which is the price we pay for predicting more positives correctly.

5. Naive Bayes

5.1. Naive Bayes without oversampling

The next classification technique we use is the Naive Bayes. It is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

Naive Bayes

The name naive is used because it assumes the features that go into the model are independent of each other. That is changing the value of one feature, does not directly influenceor change the value of any of the other features used in the algorithm represented by the following formula:

$$P(Y = k|X_1 \dots X_N) = \frac{P(X_1|Y = k) * P(X_2|Y = k) * \dots * P(X_N|Y = k) * P(Y = k)}{P(X_1) * P(X_2) * \dots * P(X_N)}$$

The confusion matrix is:

```
[[853  52]
 [117  38]]

Accuracy: 0.84
Precision: 0.42
Recall: 0.25
F1: 0.31
```

After applying the Naive Bayes algorithm, the recall is already higher than while using the logistic regression without the oversampling (0.25 instead of 0.07), accuracy stays on the same level, but precision is 40 p.p. lower.

5.2. Naive Bayes with oversampling - SMOTENC

Oversampling the train data

Following the confusion matrix analysis, we saw that the recall here was about just 25%, hence we felt the need of the introduction of SMOTENC algorithm to enable the oversampling of our train data and after that analysis, it resulted in us having a recall above 50% in the test data:

The confusion matrix is:

```
[[671 234]
 [ 63  92]]

Accuracy: 0.72
Precision: 0.28
Recall: 0.59
F1: 0.3
```

6. Random Forrest

6.1 Random Forrest without oversampling

The next algorithm is Random Forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual decision tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Running the algorithm, the winning decision tree is shown below:

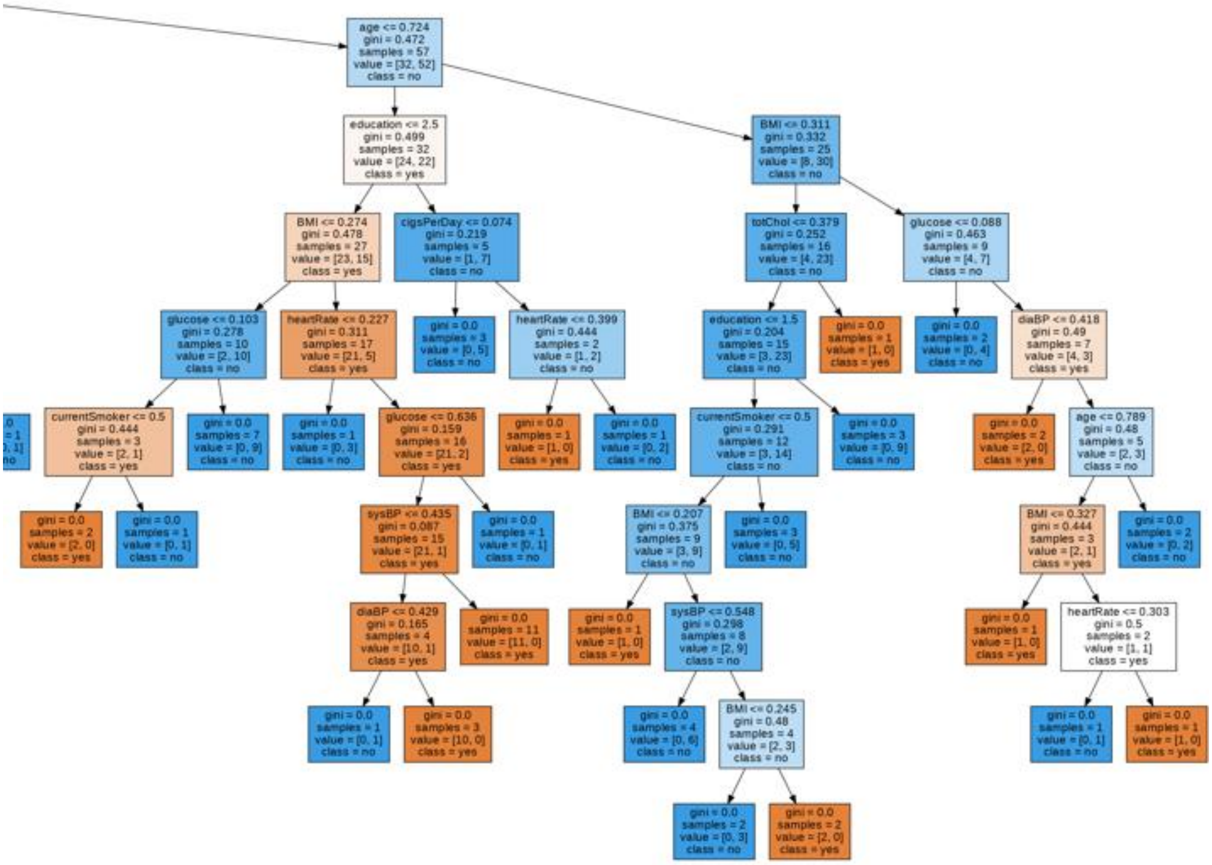


Figure 12: One part of the Decision Tree

The final tree is colossal. The root of this tree is the variable sysBP, with the Gini impurity 0.256.

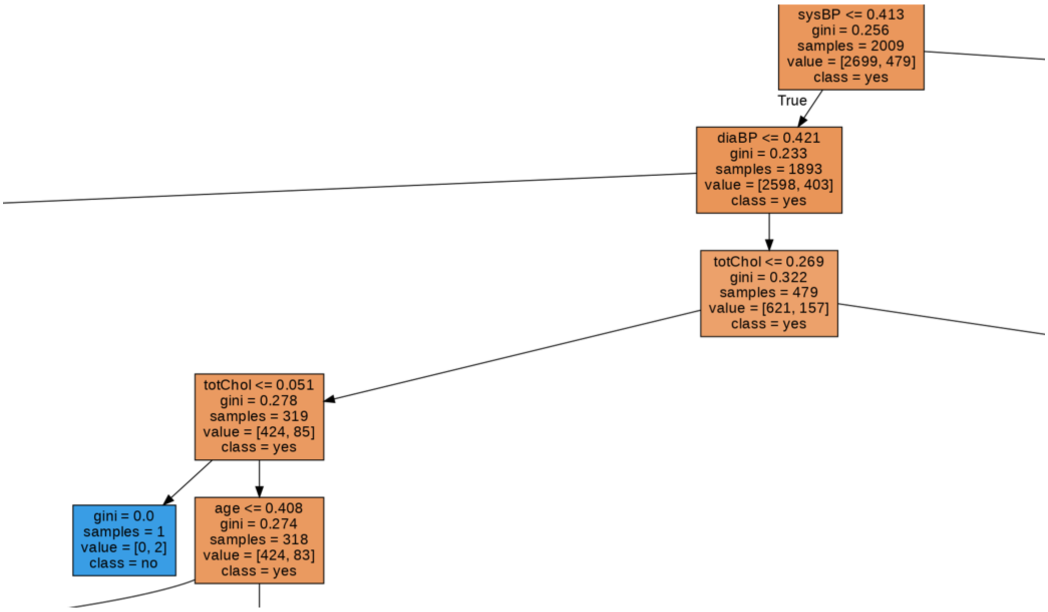


Figure 13: One part of the decision tree

Our decision tree starts with asking the question if the sysBP is smaller than 0.413 or not (reminding that the values are normalized). If smaller than the next question is about diaBP if it is smaller then 0.421 and if not the next two questions that follow are about totChol and if the totCol is less than 0.051 than the point is classified as not having a 10 year risk of coronary heart disease CHD, shortly, algorithm predicts all the points that follow this pattern of the tree with the value 0.

The confusion matrix is:

```
[[901   4]
 [147   8]]

Accuracy: 0.86
Precision: 0.67
Recall: 0.05
F1: 0.10
```

As observed, the accuracy is 0.86, precision 0.67 but the recall is 0.05. We recognize that in every model used without the oversampling technique, we have a lot more of false negatives.

6.2. Random Forest with oversampling - SMOTENC

Once again, we will introduce oversampling, SMOTENC, for this model.

The confusion matrix is:

```
[[784 121]
 [108  47]]

Accuracy: 0.78
Precision: 0.28
Recall: 0.30
F1: 0.29
```

We observe that the recall increased but the accuracy and the precision decreased. The results are comparable with Naïve Bayes with oversampling but the recall for this model is lower than the latter.

7. XGBOOST with oversampling - ADASYN Forrest

The last model we have implemented in our dataset is XGBOOST which stand for “Extreme Gradient Boostings”. XGBOOST is used for supervised learning problems and performs under the Gradient Boosting framework.

Different from the other models, we will use another oversampling technique called ADASYN. Adaptive Synthetic (ADASYN) which uses a density distribution while using the kNN algorithm to generate synthetic data.

The confusion matrix after implementing the algorithm is:

```
[[619 286]
 [ 70  85]]

Accuracy: 0.66
Precision: 0.23
Recall: 0.55
F1: 0.32
```

After applying XGBOOST with ADASYN oversampling, we observe that the accuracy is 0.66, precision 0.23 and recall 0.55. These results are comparable to the Logistic Regression 4 with the SMOTENC oversampling.

8. Conclusion

After the 4 models implemented, we must choose the winner. Among all, the logistic regression performed better on the test set. We remind that the confusion matrix and the other metrics are as below:

```
[[607 298]
 [ 61  94]]

Accuracy: 0.66
Precision: 0.24
Recall: 0.61
F1: 0.34
```

The accuracy and the precision of this model is lower than of the Naïve Bayes and Random Forest with oversampling. These metrics are almost the same for the XGBOOST algorithm with oversampling. However, the Recall for this logistic regression is the highest among all, resulting at 0.6.

The reason why we are interested in a high Recall, is that we prefer that the positive cases to be predicted correctly within our model. The sacrifice of choosing this model comes with a low accuracy and precision. Despite that, this model results in a lot more false positives points than other models. We argue that this is not a problem, as in case of predicting that the person will have a 10 year risk of coronary heart disease CHD while not true, the person is subjected to other medical tests, where the result might change.

The ROC curve, another performance measurement, of this model is as below:

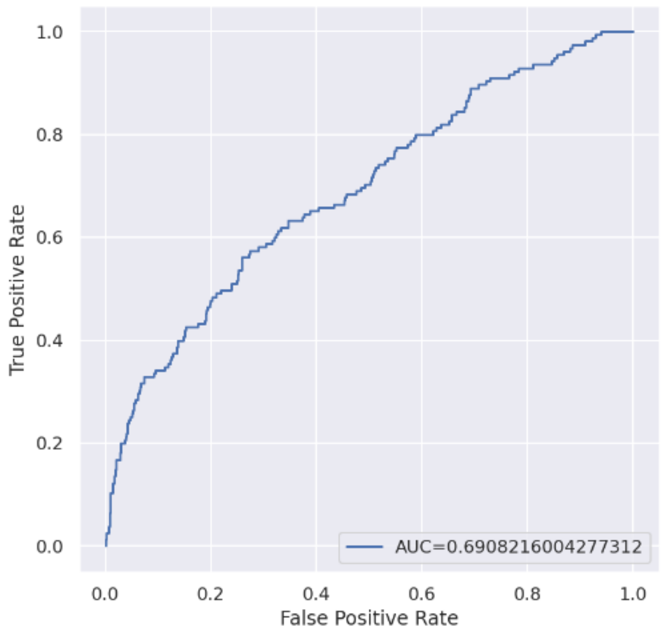


Figure 14: ROC Curve for the Logistic Regression with oversampling

The area under the curve is 0.69, the AUC ranges between 0.5 (worst) to 1 (best).

9. Application to business

Medical data is a valuable source of information. The models we introduced in the previous chapters can be further developed and become an instrument for banks and insurance companies to reduce the risk of fraud. These companies typically face the hidden information problem – clients try to hide information about their medical records and companies, therefore, misclassify this client as a low risk.

Even though the models analyzed tend to show good enough results in terms of predicting the heart disease of a patient (e.g. the "champion" model – Logistic regression with oversampling, which provides the recall and accuracy of more than 60%), there are other problems connected with the analysis, such as loosing profit from clients who were classified as high risk and are actually low risk (for the last logistic regression with oversampling, 298 points were classified as false positive).

Nevertheless, insurance companies tend to use medical data a lot in order to reduce the risks connected with paying out the claims. They keep investing in the artificial intelligence and machine learning technologies to precisely analyze the medical records. Such health insurers as Cigna Corp. and UnitedHealthcare use the electronic medical records, information from surveys and data from claims documents to determine the risk of health problem for the client (including calculating the probability of having the coronary heart disease). UnitedHealthcare, for instance, invests 5 bln $ yearly into data and technologies. These companies invest into programs which help people to improve their health. Clients who have been classified as high risk – due to the rise in blood sugar or cholesterol level (which as we have analyzed are the factors which significantly affect the risk of having a heart disease in the nearest 10 years) are being contacted via phone calls by nurses and suggested to join the health program. Cigna Corp. reports that this initiative has led to 10% increase in participation in medical programs.

Health and life insurance companies typically require clients to provide medical record and/or to do medical testing prior to purchasing the policy. They also tend to check other medical data available, such as family medical records. After it, the applicants with diabetes, high cholesterol or higher blood pressure will be assigned the higher premiums since they will be qualified as the high-risk group.

In the USA the insurance companies are a part of Medical Information Bureau (MIB), which unites over 600 life and health insurers and provides medical information about clients gained from the precious applications. Insurance companies can also get into collaboration with prescription drug databases (such as MedPoint) to get the information about client's recent purchase on drugs in the drug stores from which it will be possible to estimate some assumptions about client's health status.

Evidence from Singapore shows that the insurance industry is quite small and even though all the medical data is stored into the National Electronic Health Records (NEHR) and insurance companies don't have a direct access to it, the crosscheck with the other insurers makes it almost impossible for clients to hide their medical information.

It's important to note that in general, the medical data is protected and not an open source for the insurance companies. In most cases insurance companies get access to medical data only in specific cases connected with paying out claims or determining coverage eligibility. When the specific medical characteristics are not observed, the insurance companies use the data which can be addressed as instrumental variables. For example, changing a name for a woman can be considered as a sign of depression after changing the medical status (e.g., divorce) and increase the risk score. At the same time, buying plus size clothes is a symptom of obesity and health problems connected to it.

Another type of companies who can benefit from medical information are banks. Classifying the clients as high-risk groups can allow the banks to refuse such clients on long term loans or to increase the loan/mortgage interest rate.

References

Anderson, K.M.; Wilson, P.W.; Odell, P.M.; Kannel, W.B. An updated coronary risk profile. A statement for health professionals. Circulation 1991, 83, 356 to 362.

Wilson, P.W.F.; D Agostino, R.B.; Levy, D.; Belanger, A.M.; Silbershatz, H.; Kannel, W.B. Prediction of coronary heart disease using risk factor categories. Circulation 1998, 97, 1837 to 1847.

D Agostino, R.B.S.; Grundy, S.; Sullivan, L.M.; Wilson, P.; CHD Risk Prediction Group. Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. JAMA 2001, 286, 180 to 187.

D Agostino, R.B.; Vasan, R.S.; Pencina, M.J.; Wolf, P.A.; Cobain, M.; Massaro, J.M.; Kannel, W.B. General cardiovascular risk Profile for use in primary care: The Framingham heart study. Circulation 2008, 117, 743 to 753.

Hense, H._W.; Schulte, H.; LÃ¶wel, H.; Assmann, G.; Keil, U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany ;results from the MONICA Augsburg and the PROCAM cohorts. Eur. Heart J. 2003, 24, 937 to 945.

Brindle, P.; Jonathan, E.; Lampe, F.; Walker, M.; Whincup, P.; Fahey, T.; Ebrahim, S. Predictive accuracy of the Framingham coronary risk score in British men: Prospective cohort study. BMJ 2003, 327, 1267.

Liu, J.; Hong, Y.; D Agostino, S.R.B.; Wu, Z.; Wang, W.; Sun, J.; Wilson, P.W.F.; Kannel, W.B.; Zhao, D. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. JAMA 2004, 291, 2591 to 2599.

Ridker, P.M.; Buring, J.E.; Rifai, N.; Cook, N.R. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score. JAMA 2007, 297, 611 to 619.

Lloyd-Jones, D.M.; Leip, E.P.; Larson, M.; D Agostino, R.B.; Beiser, A.; Wilson, P.W.; Wolf, P.A.; Levy, D. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 2006, 113, 791 to 798.

Pencina, M.J.; D Agostino, R.B.; Larson, M.G.; Massaro, J.M.; Vasan, R.S. Predicting the 30-year risk of cardiovascular disease: The Framingham Heart Study. Circulation 2009, 119, 3078 to 3084.

Can Health Insurance Companies Access Medical Records? 18 August 2022. https://www.helpadvisor.com/insurance/insurance-companies-and-medical-records.

Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates. 17 July 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Journal, The Wall Street. Health Insurers Have the Data. Will Patients Listen? 5 July 2022. https://www.wsj.com/articles/health-insurers-have-the-data-will-patients-listen-11657013401.

The importance of medical records for insurance companies. 16 April 2021. https://www.recordrs.com/blog/the-importance-of-medical-records-for-insurance-companies/.

Understanding the maths behind the Gini impurity method for decision tree split. 21 December 2022. https://analyticsindiamag.com/understanding-the-maths-behind-the-gini-impurity-method-for-decision-tree-split/.