Machine Learning Assignment1
Classification

Lovejeet Singh Parihar
(IMT2019048)

3 October 2021

1 Evaluation metrics used

The metrics used for evaluation of the models are Accuracy, precision, recall and f1 score. We find out that the dataset is imbalanced and just predicting all 0s provides us an accuracy of about 0.85. Hence we will focus on other metrics such as precision, recall and f1score and try to improve them.

2 Data Preprocessing

2.1 Imports and visualisation of structure

We import the numpy, pandas and matplotlib libraries.
The raw data has 4238 rows and 16 columns.
The ratio of ones to the total number of values in the dataset = 0.15195847. Hence we can say that the dataset is imbalanced and there are less 1's compared to 0's.

2.2 Handling Null values

- Null values of cigsPerDay

  Get the median cigerattes smoked by a smoker. For smoker nans, use this median value. For non smoker nans, use 0.

- Null values of education

  Replace by median value of education.

- Null values of BMI

  Replace by mean value of BMI.

- Null values of totChol

  Replace by mean value of totChol.

- Null values of BPMeds

  From the data, we note that:

  1. Everyone who takes BPMeds has prevalent Hypertension
  2. Everyone who has prevalent Hypertension may not take BPMeds
  3. Probability that someone suffering with prevalent Hypertension takes BPMeds is about 0.0942

  So we will generate a random array (of 0's and 1's) with this probability of 1s and assign it to nulls in BP Meds.

- Null values of glucose

  Remove all the rows of glucose nans as it gave a better answer than trying other options like replacing with mean/median etc.

- Null values of heartRate

  Remove all the rows of heartRate nans, as there is just one nan and you will remove atmost one row.

2.3 Checking correlation

As expected, the columns cigsPerDay and current Smoker are highly correlated we can try removing the currentSmoker coumn as cigsPerDay naturally cover the fact that the person is a current Smoker or not.

We have the diaBP and sysBP highly correlated. As per my research, the sysBP is more important from the perspective of Heart disease risk, we can try removing diaBP.

But as I have tried, removing these columns didn't increase the scores instead it decreased them a bit, I decided not to remove them.

These highly correlated columns like sysBP and diaBP can be modelled better using a multivariate gaussian bayes classifier rather than a gaussian naive bayes classifier. This is because in naive bayes we assume the columns to be independent and we don't capture the correlation between the columns.

2.4 Normalisation

The normalisation for features is done here is done by subtracting the mean and dividing by the standard deviation.

It helps easy and fast convergence of the gradient descent algorithm.

2.5 Splitting into test and training set

The split was according to the parameters mentioned in the assignment pdf, i.e.

```
X_train , X_test , y_train , y_test=
train_test_split(X,y, test_size=.20,  random_state=5)
```

2.6 Duplicating rows for handling imbalance

As the dataset is imbalanced, we make the number of ones approximately 2.75 times the original amount by duplicating the rows containing ones and adding them at the end. Much higher values cause recall to go up and precision to go down, contributing to a lower f1 score, and much lower values cause precision to go up and recall to go down, again contributing to a lower f1 score.

2.7 Polynomial Columns

I also tried to add some more polynomial features of the second degree, but it did not improve the metrics much so I reverted back.

2.8 Outlier Handling

I tried removing the 'outliers', but it worsened the answer. It removed some rare events like glucose above 400, but they were very much possible and we should not ignore them as they can be key to classification of the final variable.

3 Logistic Regression using Gradient Descent

- For all features left after cleaning

  Accuracy : 0.8116883116883117

  Recall : 0.3969465648854962

  Precision : 0.4406779661016949

  F1 Score : 0.4176706827309237

- For age :

  Accuracy : 0.7896103896103897

  Recall : 0.21374045801526717

  Precision : 0.3218390804597701

  F1 Score : 0.2568807339449541

- For cigsPerDay :
  Accuracy : 0.8298701298701299
  Recall : 0
  Precision : 0
  F1 Score : 0

- For totChol :
  Accuracy : 0.8298701298701299
  Recall : 0.022900763358778626
  Precision : 0.5
  F1 Score : 0.043795620437956206

- For BMI :
  Accuracy : 0.8298701298701299
  Recall : 0.007633587786259542
  Precision : 0.5
  F1 Score : 0.015037593984962405

- For heartRate :
  Accuracy : 0.8298701298701299
  Recall : 0
  Precision : 0
  F1 Score : 0

- For sysBP :
  Accuracy : 0.8103896103896104
  Recall : 0.17557251908396945
  Precision : 0.3770491803278688
  F1 Score : 0.23958333333333333

- For glucose :
  Accuracy : 0.8324675324675325
  Recall : 0.05343511450381679
  Precision : 0.5833333333333334
  F1 Score : 0.09790209790209789

- For the combination of age, cigsPerDay, totChol, BMI, heartRate, sysBP and glucose
  Accuracy : 0.8194805194805195
  Recall : 0.3511450381679389
  Precision : 0.46
  F1 Score : 0.3982683982683982

4 Logistic Regression Using Newton's method in Gradient Descent

- For all features left after cleaning
  Accuracy : 0.8116883116883117
  Recall : 0.3969465648854962
  Precision : 0.4406779661016949
  F1 Score : 0.4176706827309237

- For age :
  Accuracy : 0.7896103896103897
  Recall : 0.21374045801526717
  Precision : 0.3218390804597701
  F1 Score : 0.2568807339449541

- For cigsPerDay :
  Accuracy : 0.8298701298701299
  Recall : 0
  Precision : 0
  F1 Score : 0

- For totChol :
  Accuracy : 0.8298701298701299
  Recall : 0.022900763358778626
  Precision : 0.5
  F1 Score : 0.043795620437956206

- For BMI :
  Accuracy : 0.8298701298701299
  Recall : 0.007633587786259542
  Precision : 0.5
  F1 Score : 0.015037593984962405

- For heartRate :
  Accuracy : 0.8298701298701299
  Recall : 0
  Precision : 0
  F1 Score : 0

- For sysBP :
  Accuracy : 0.8103896103896104
  Recall : 0.17557251908396945
  Precision : 0.3770491803278688
  F1 Score : 0.23958333333333333

- For glucose :
  Accuracy : 0.8324675324675325
  Recall : 0.05343511450381679
  Precision : 0.5833333333333334
  F1 Score : 0.09790209790209789

- For the combination of age, cigsPerDay, totChol, BMI, heartRate, sysBP and glucose
  Accuracy : 0.8194805194805195
  Recall : 0.3511450381679389
  Precision : 0.46
  F1 Score : 0.3982683982683982

5 Naive Bayes Classifier

- For all features left after cleaning
  Accuracy : 0.7935064935064935
  Recall : 0.24427480916030533
  Precision : 0.34782608695652173
  F1 Score : 0.28699551569506726

- For age :
  Accuracy : 0.7766233766233767
  Recall : 0.24427480916030533
  Precision : 0.3047619047619048
  F1 Score : 0.2711864406779661

- For cigsPerDay :
  Accuracy : 0.8155844155844156
  Recall : 0.07633587786259542
  Precision : 0.3225806451612903
  F1 Score : 0.1234567901234568

- For totChol :
  Accuracy : 0.8116883116883117
  Recall : 0.04580152671755725
  Precision : 0.23076923076923078
  F1 Score : 0.07643312101910829

- For BMI :
  Accuracy : 0.8051948051948052
  Recall : 0.04580152671755725
  Precision : 0.1935483870967742
  F1 Score : 0.07407407407407407

- For heartRate :
  Accuracy : 0.8298701298701299
  Recall : 0
  Precision : 0
  F1 Score : 0

- For sysBP :
  Accuracy : 0.8
  Recall : 0.22137404580152673
  Precision : 0.35802469135802467
  F1 Score : 0.27358490566037735

- For glucose :
  Accuracy : 0.8246753246753247
  Recall : 0.07633587786259542
  Precision : 0.4166666666666667
  F1 Score : 0.12903225806451613

- For the combination of age, cigsPerDay, totChol, BMI, heartRate, sysBP and glucose

  Accuracy : 0.8194805194805195

  Recall : 0.3511450381679389

  Precision : 0.46

  F1 Score : 0.3982683982683982

6 Some observations

- Normalisation caused the gradient descent to converge in lesser number of iterations.
- The number of iterations using Newton's method for gradient descent is significantly lower than when we don't use it.
- For this case, the overall answer that we get by logistic regression is better than that of Naive Bayes.
- On their individual capacity, Age and sysBP are two of the important factors predicting whether the person is likely to get heart disease in next 10 years.