Jiajun Li

JL11895@nyu.edu

Homework 3 Machine Learning

Idea of Solving question 1:

Firs of all, the dataset should be cleaned. All the rows containing NaN is removed from the dataset and all the non-dummy variables are normalized to improve the model performance.

Secondly, to solve class imbalance problem. The dataset is resample for 100 times. I used to RandomUnderSampler(random_state = i) to downsample the dataset because downsampling is more efficient for computation and reduce the risk of overfitting. The random_state is synchronized with the index of the iteration.

Finally, Calculating the AUC score of this model. For each time of the iteration after downsampling. I split the dataset into train and test set and then I used the training set to fit the LogisticRegressionCV() model because this method support K-hold cross validation to improve the beta coefficients of the model. I used 5-folds to tune the beta coefficients. And then the AUC can be calculated by the roc_auc_score. Adding all the 100 AUC scores and take the average to generate the final AUC score. I am doing in this way to reduce the variance in the AUC score, which is possibly influenced by how the dataset is resampled.

To find the best predictor, my strategy is implement a for loop which iterates 21 times and for each time, one predictor is removed. The AUC score after removing each predictor is calculated. The removed predictor that causes the AUC score to drop the most should be the best predictor. Of course, the dataset is resampled for 10 times for each of the 21 iterations and the final AUC is the average of the all the 10 AUC scores.

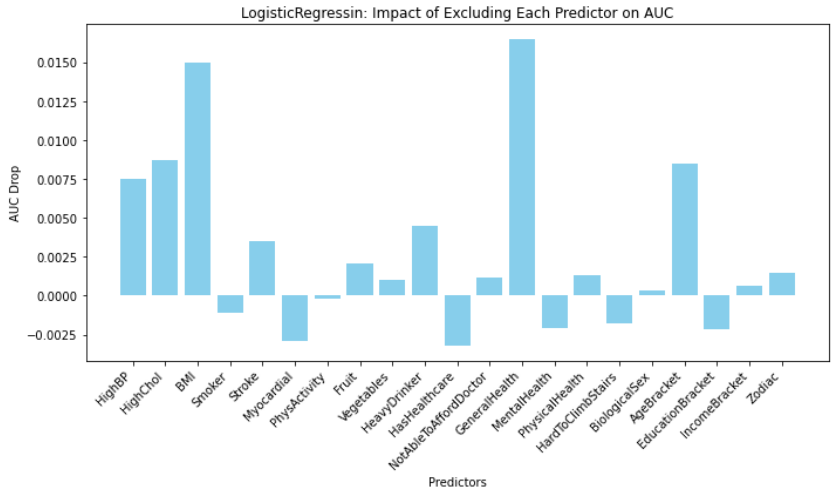
Data and Graph:

Here is the AUC score of full predictors and all 21 iterations which one of the predictors is not included. The drop in AUC score is illustrated in a bar chart:

The AUC score for LogisticRegression Model is 0.8189752862617599

The best predictor is GeneralHealth: AUC Score after excluding GeneralHealth: 0.8025259123249906 The AUC drops by: 0.016449373936769263

```
LogisticRegresssion AUC Score: 0.8189752862617599
LogisticRegression AUC Score after excluding HighBP: 0.8114837683856655
LogisticRegression: The AUC drops by: 0.007491517876094367
LogisticRegression AUC Score after excluding HighChol: 0.8102404940573971
LogisticRegression: The AUC drops by: 0.008734792204362796
LogisticRegression AUC Score after excluding BMI: 0.8040088515290635
LogisticRegression: The AUC drops by: 0.014966434732696321
LogisticRegression AUC Score after excluding Smoker: 0.8201004619803518
LogisticRegression: The AUC drops by: -0.001125175718591942
LogisticRegression AUC Score after excluding Stroke: 0.8154817301802897
LogisticRegression: The AUC drops by: 0.0034935560814701727
LogisticRegression AUC Score after excluding Myocardial: 0.8218995716920874
LogisticRegression: The AUC drops by: -0.0029242854303275223
LogisticRegression AUC Score after excluding PhysActivity: 0.8192006714351976
LogisticRegression: The AUC drops by: -0.000225385173437731
LogisticRegression AUC Score after excluding Fruit: 0.8168737955999514
LogisticRegression: The AUC drops by: 0.002101490661808425
LogisticRegression AUC Score after excluding Vegetables: 0.8179925592803248
LogisticRegression: The AUC drops by: 0.0009827269814350892
LogisticRegression AUC Score after excluding HeavyDrinker: 0.8145061054953964
LogisticRegression: The AUC drops by: 0.00446918076636349
LogisticRegression AUC Score after excluding HasHealthcare: 0.8221619872545103
LogisticRegression: The AUC drops by: -0.003186700992750424
LogisticRegression AUC Score after excluding NotAbleToAffordDoctor: 0.8177773527267254
LogisticRegression: The AUC drops by: 0.001197933535034501
LogisticRegression AUC Score after excluding GeneralHealth: 0.8025259123249906
LogisticRegression: The AUC drops by: 0.016449373936769263
LogisticRegression AUC Score after excluding MentalHealth: 0.8210402396588277
LogisticRegression: The AUC drops by: -0.0020649533970678036
LogisticRegression AUC Score after excluding PhysicalHealth: 0.8176562884787775
LogisticRegression: The AUC drops by: 0.0013189977829823896
LogisticRegression AUC Score after excluding HardToClimbStairs: 0.8207804344664293
LogisticRegression: The AUC drops by: -0.0018051482046694822
LogisticRegression AUC Score after excluding BiologicalSex: 0.8186355374423249
LogisticRegression: The AUC drops by: 0.0003397488194349929
LogisticRegression AUC Score after excluding AgeBracket: 0.8104986094020796
LogisticRegression: The AUC drops by: 0.008476676859680232
LogisticRegression AUC Score after excluding EducationBracket: 0.8211050853414452
LogisticRegression: The AUC drops by: -0.002129799079685357
LogisticRegression AUC Score after excluding IncomeBracket: 0.8183022683439576
LogisticRegression: The AUC drops by: 0.0006730179178022766
LogisticRegression AUC Score after excluding Zodiac: 0.8174983010582351
LogisticRegression: The AUC drops by: 0.0014769852035247943
```



We can see that the AUC is around 0.8, which means that the Logistic Regression model has a great performance in distinguishing the different class. And since removing predictor GeneralHealth drops the AUC value the most, we should conclude that GeneralHealth is the best

predictor. The removal of the predictor that causes the AUC score to increase means that they are redundant in building the model.

Idea of Solving q2:

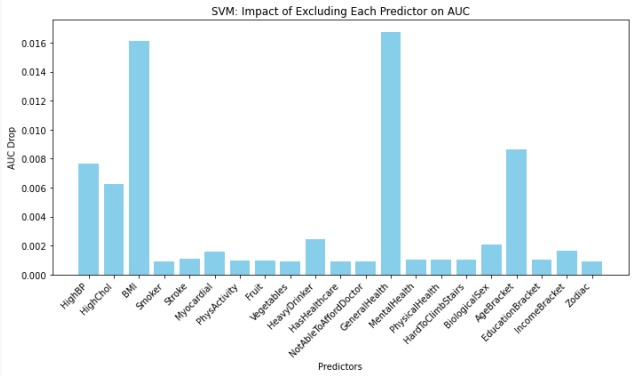
The general idea of question 2 is similar to that of question 1. The first part is to calculate the average of the 100 AUC scores (all of the predictors). The second part is to exclude each of the predictors to see which one of them drops the AUC score the most. The difference is that I implemented a hyperparameter tuning process.

Tunning Process:

First of all, I created a array of 100 C values ranging from 0.001 to 100. After initialize the LinearSVC model who uses a linear Kernel, I used gridsearch to optimize the C value and find the best model. Then since the LinearSVC has no built-in methods to calculate the AUC value, calibratedClassifierCV is used instead.

Data and Graph:

```
Linear SVM AUC Score: 0.8185631975963753
SVM: AUC Score after excluding HighBP: 0.8109205870562983
SVM: The AUC drops by: 0.007642610540076977
SVM: AUC Score after excluding HighChol: 0.8122944754955939
SVM: The AUC drops by: 0.006268722100781443
SVM: AUC Score after excluding BMI: 0.8024495481067049
SVM: The AUC drops by: 0.016113649489670423
SVM: AUC Score after excluding Smoker: 0.8176428190583905
SVM: The AUC drops by: 0.0009203785379847673
SVM: AUC Score after excluding Stroke: 0.8174952926846037
SVM: The AUC drops by: 0.0010679049117715556
SVM: AUC Score after excluding Myocardial: 0.8170052310659079
SVM: The AUC drops by: 0.0015579665304673584
SVM: AUC Score after excluding PhysActivity: 0.817595176025522
SVM: The AUC drops by: 0.0009680215708532947
SVM: AUC Score after excluding Fruit: 0.8176156818828592
SVM: The AUC drops by: 0.0009475157135161449
SVM: AUC Score after excluding Vegetables: 0.8176444146303921
SVM: The AUC drops by: 0.0009187829659832447
SVM: AUC Score after excluding HeavyDrinker: 0.8161264546664955
SVM: The AUC drops by: 0.0024367429298798315
SVM: AUC Score after excluding HasHealthcare: 0.8176501694512586
SVM: The AUC drops by: 0.0009130281451167344
SVM: AUC Score after excluding NotAbleToAffordDoctor: 0.8176394980629322
SVM: The AUC drops by: 0.0009236995334430675
SVM: AUC Score after excluding GeneralHealth: 0.8018329133365067
SVM: The AUC drops by: 0.016730284259868555
SVM: AUC Score after excluding MentalHealth: 0.8175215586991685
SVM: The AUC drops by: 0.0010416388972067647
SVM: AUC Score after excluding PhysicalHealth: 0.8175217761558876
SVM: The AUC drops by: 0.0010414214404876887
SVM: AUC Score after excluding HardToClimbStairs: 0.8175111848129791
SVM: The AUC drops by: 0.0010520127833961723
SVM: AUC Score after excluding BiologicalSex: 0.8165007550461943
SVM: The AUC drops by: 0.002062442550180954
SVM: AUC Score after excluding AgeBracket: 0.809901682258521
SVM: The AUC drops by: 0.008661515337854309
SVM: AUC Score after excluding EducationBracket: 0.8175189167556735
SVM: The AUC drops by: 0.00104428084070185
SVM: AUC Score after excluding IncomeBracket: 0.8169054362196464
SVM: The AUC drops by: 0.0016577613767289057
SVM: AUC Score after excluding Zodiac: 0.817644383946315
SVM: The AUC drops by: 0.0009188136500603017
```



Linear SVM AUC Score: 0.8185631975963753

SVM: AUC Score after excluding GeneralHealth: 0.8018329133365067

SVM: The AUC drops by: 0.016730284259868555

The best predictor is indeed GeneralHealth.

From the result, we can see that SVM has a great performance in distinguishing the different classes and since the removal of GeneralHealth caused the AUC score to drop the most, GeneralHealth should be the best predictor. Also from the graph, BMI and AgeBracket can also be very robust predictors.

Idea of Solving q3

The basic structure of solving q3 is also quite similar to that of q1. The first part is to build the DecisionTree model with all of the 21 predictors included and the second part is to find out the best predictor by finding which one of the excluded predictors drops the AUC the most.

However, the difference is that the tunning process is applied to improve the performance of the decision tree model. By researching, the performance of the decision tree model is strongly related to the maximum depths, criterion used to decide the root and split the branches, minimum samples used to determine the split and leaf. These factors influence the tree greatly and thus needed to be optimized. At the same time, the tuning process can be quite computationally expensive so in the first part, the iteration is reduced from 100 to 50 in order to improve computation efficiency. The parameters to be tuned are: max_depth, minimum sample split, minimum sample leaf and criterion which is gini impurity in this case .

Notice: max_depth is critical to prevent the model from overfitting and underfitting. Minimum sample split also contributes to preventing the model from being too complex and thus result in overfitting. Minimum sample leaf determines the minimum number of leaves that a model must process, which makes the model smoother.

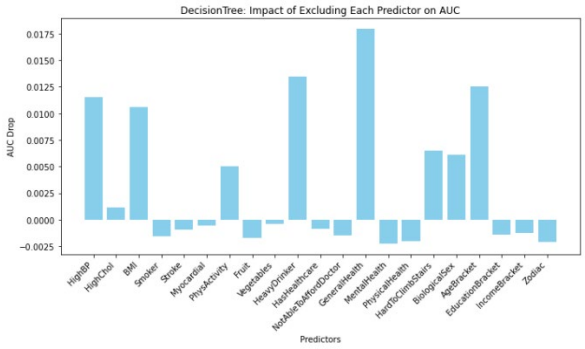
And the RandomizedSearchCV method is used to find the best model so we can calculate the AUC score.

Data and Graph:

AUC Score for DecisioinTree : 0.7962304608800831

DecisionTree: AUC Score after excluding GeneralHealth: 0.7783034577299315

DecisionTree: The AUC drops by: 0.01792700315015161




```
q3:
AUC Score for DecisioinTree : 0.7962304608800831
DecisionTree: AUC Score after excluding HighBP: 0.7846648205287895
DecisionTree: The AUC drops by: 0.011565640351293593
DecisionTree: AUC Score after excluding HighChol: 0.7950803476785188
DecisionTree: The AUC drops by: 0.001150113201564329
DecisionTree: AUC Score after excluding BMI: 0.7856551993630306
DecisionTree: The AUC drops by: 0.010575261517052459
DecisionTree: AUC Score after excluding Smoker: 0.7977918622039116
DecisionTree: The AUC drops by: -0.0015614013238285374
DecisionTree: AUC Score after excluding Stroke: 0.7971542759902159
DecisionTree: The AUC drops by: -0.0009238151101328418
DecisionTree: AUC Score after excluding Myocardial: 0.7967838782693473
DecisionTree: The AUC drops by: -0.0005534173892641769
DecisionTree: AUC Score after excluding PhysActivity: 0.7911829896335758
DecisionTree: The AUC drops by: 0.005047471246507285
DecisionTree: AUC Score after excluding Fruit: 0.7979457450732912
DecisionTree: The AUC drops by: -0.0017152841932080998
DecisionTree: AUC Score after excluding Vegetables: 0.7966447726735473
DecisionTree: The AUC drops by: -0.0004143117934641838
DecisionTree: AUC Score after excluding HeavyDrinker: 0.7827457814352565
DecisionTree: The AUC drops by: 0.013484679444826564
DecisionTree: AUC Score after excluding HasHealthcare: 0.797101592764186
DecisionTree: The AUC drops by: -0.000871131884102927
DecisionTree: AUC Score after excluding NotAbleToAffordDoctor: 0.7976862275997011
DecisionTree: The AUC drops by: -0.0014557667196180013
DecisionTree: AUC Score after excluding GeneralHealth: 0.7783034577299315
DecisionTree: The AUC drops by: 0.01792700315015161
DecisionTree: AUC Score after excluding MentalHealth: 0.798466457416678
DecisionTree: The AUC drops by: -0.0022359965365948975
DecisionTree: AUC Score after excluding PhysicalHealth: 0.7982656710451541
DecisionTree: The AUC drops by: -0.0020352101650710 22
DecisionTree: AUC Score after excluding HardToClimbStairs: 0.7897004702232513
DecisionTree: The AUC drops by: 0.006529990656831797
DecisionTree: AUC Score after excluding BiologicalSex: 0.7900948295400358
DecisionTree: The AUC drops by: 0.006135631340047243
DecisionTree: AUC Score after excluding AgeBracket: 0.7836733628601908
DecisionTree: The AUC drops by: 0.012557098019892243
DecisionTree: AUC Score after excluding EducationBracket: 0.7976492306074433
DecisionTree: The AUC drops by: -0.0014187697273602 14
DecisionTree: AUC Score after excluding IncomeBracket: 0.7974510145832345
DecisionTree: The AUC drops by: -0.0012205537031514 39
DecisionTree: AUC Score after excluding Zodiac: 0.7983270463141988
DecisionTree: The AUC drops by: -0.0020965854341157275
```

As one can see, the Decision Tree model also has a great performance even though not as good as the first two and again the best predictor is General Health.

Idea of Solving q4:

The process of solving q4 is also the same as solving the first three problems. The goal is to find the AUC score of the model including all 21 predictors and then find the best predictor. The final AUC score is the average of all the AUC scores in each iteration.

Similarly in question 4, the hyperparameters should be tuned are: number of trees, maximum depth of each tree, minimum number of samples required to split an internal node, Minimum number of samples required to be at a leaf node, whether bootstrap is used, and the criterion used to decide the root and split the branch. Generally, more trees generated will increase the randomforest model's performance. The implementation of bootstrap will assign more accuracy (reduce bias and variance). Again the criterion here is gini importance. The tuning process is achieved by RandomizedSearchCV to find the best estimator and reduce the computation cost and to further reduce computation cost, the iteration of the first part is reduced from 100 to 50. So the AUC score is the average of 50 AUCs from models built on different data samples.

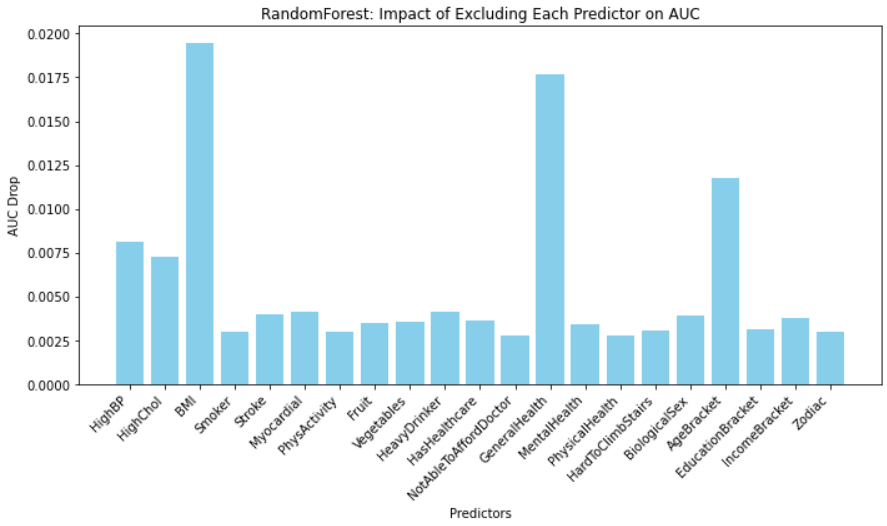
Data and Graph:

AUC Score for RandomForest : 0.8221979892822211

RandomForest: AUC Score after excluding BMI: 0.8027383172889555

RandomForest: The AUC drops by: 0.019459671993265593

```
AUC Score for RandomForest : 0.8221979892822211
RandomForest: AUC Score after excluding HighBP: 0.8141140430374152
RandomForest: The AUC drops by: 0.008083946244805906
RandomForest: AUC Score after excluding HighChol: 0.8149164183087103
RandomForest: The AUC drops by: 0.007281570973510787
RandomForest: AUC Score after excluding BMI: 0.8027383172889555
RandomForest: The AUC drops by: 0.019459671993265593
RandomForest: AUC Score after excluding Smoker: 0.8191953484148924
RandomForest: The AUC drops by: 0.0030026408673287097
RandomForest: AUC Score after excluding Stroke: 0.8182345499202228
RandomForest: The AUC drops by: 0.003963439361998278
RandomForest: AUC Score after excluding Myocardial: 0.8180639375584546
RandomForest: The AUC drops by: 0.004134051723766463
RandomForest: AUC Score after excluding PhysActivity: 0.819234801912093
RandomForest: The AUC drops by: 0.0029631873701281286
RandomForest: AUC Score after excluding Fruit: 0.8187269270749864
RandomForest: The AUC drops by: 0.0034710622072346986
RandomForest: AUC Score after excluding Vegetables: 0.818655108547057
RandomForest: The AUC drops by: 0.0035428807351640668
RandomForest: AUC Score after excluding HeavyDrinker: 0.8180688425637987
RandomForest: The AUC drops by: 0.00412914671842235
RandomForest: AUC Score after excluding HasHealthcare: 0.81857408924301
RandomForest: The AUC drops by: 0.003623900039211092
RandomForest: AUC Score after excluding NotAbleToAffordDoctor: 0.819464976958882
RandomForest: The AUC drops by: 0.0027330123233391124
RandomForest: AUC Score after excluding GeneralHealth: 0.804521987128928
RandomForest: The AUC drops by: 0.017676002153293124
RandomForest: AUC Score after excluding MentalHealth: 0.8188091470603049
RandomForest: The AUC drops by: 0.003388842221916155
RandomForest: AUC Score after excluding PhysicalHealth: 0.8194533703732534
RandomForest: The AUC drops by: 0.002744618908967711
RandomForest: AUC Score after excluding HardToClimbStairs: 0.8191777517638075
RandomForest: The AUC drops by: 0.0030202375184136265
RandomForest: AUC Score after excluding BiologicalSex: 0.8182952999456126
RandomForest: The AUC drops by: 0.0039026893366085336
RandomForest: AUC Score after excluding AgeBracket: 0.8104779443433001
RandomForest: The AUC drops by: 0.011720044938920982
RandomForest: AUC Score after excluding EducationBracket: 0.8190803364899482
RandomForest: The AUC drops by: 0.0031176527922729402
RandomForest: AUC Score after excluding IncomeBracket: 0.8184009376627006
RandomForest: The AUC drops by: 0.003797051619520486
RandomForest: AUC Score after excluding Zodiac: 0.8192501439505671
RandomForest: The AUC drops by: 0.002947845331654042
```

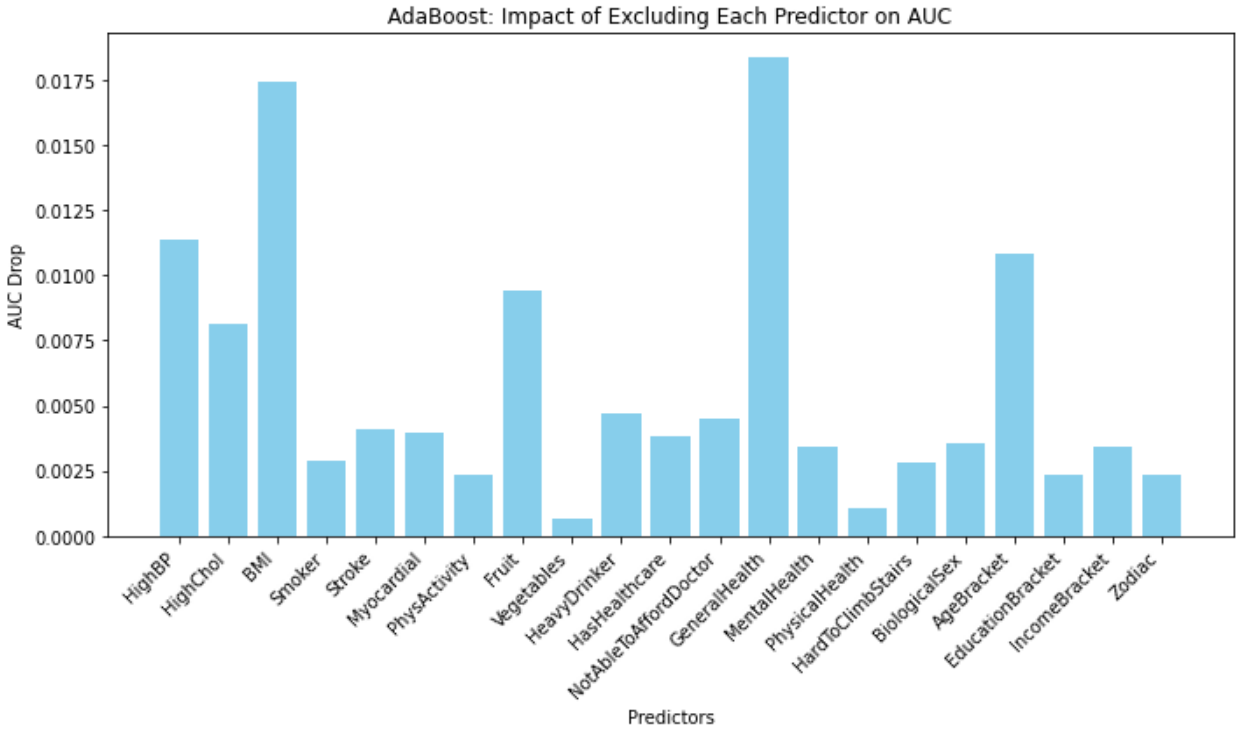


We can see that after ensemble methods, the AUC score of the tree decision model indeed increases and the now the best estimator is BMI rather than GeneralHealth. However, this is acceptable because in the previous models, BMI is within the top three best predictors even though it is not the best.

Idea of solving q5:

The idea of solving q5 follows the same pattern for the same reasons as the previous four models. The hyperparameters: max_depth, number of trees, learning rate, and algorithm are tuned to improve the model's performance.

Data and Graph:




```
q5:
AUC Score for adaBoost : 0.8236114092185303
AdaBoost: AUC Score after excluding HighBP: 0.8122492827419137
AdaBoost: The AUC drops by: 0.011362126476616607
AdaBoost: AUC Score after excluding HighChol: 0.8154489871573085
AdaBoost: The AUC drops by: 0.008162422061221752
AdaBoost: AUC Score after excluding BMI: 0.8061575862862702
AdaBoost: The AUC drops by: 0.017453822932260055
AdaBoost: AUC Score after excluding Smoker: 0.8207235132802062
AdaBoost: The AUC drops by: 0.0028878959383240588
AdaBoost: AUC Score after excluding Stroke: 0.8195049952209772
AdaBoost: The AUC drops by: 0.004106413997553071
AdaBoost: AUC Score after excluding Myocardial: 0.8196206163804927
AdaBoost: The AUC drops by: 0.0039907928380376045
AdaBoost: AUC Score after excluding PhysActivity: 0.8212772230131253
AdaBoost: The AUC drops by: 0.002334186205404931
AdaBoost: AUC Score after excluding Fruit: 0.8142224111926548
AdaBoost: The AUC drops by: 0.00938899802587545
AdaBoost: AUC Score after excluding Vegetables: 0.8229451116027462
AdaBoost: The AUC drops by: 0.0006662976157840728
AdaBoost: AUC Score after excluding HeavyDrinker: 0.818893715044555
AdaBoost: The AUC drops by: 0.004717694173975273
AdaBoost: AUC Score after excluding HasHealthcare: 0.8198068464927302
AdaBoost: The AUC drops by: 0.003804562725800098
AdaBoost: AUC Score after excluding NotAbleToAffordDoctor: 0.8190876295169327
AdaBoost: The AUC drops by: 0.004523779701597563
AdaBoost: AUC Score after excluding GeneralHealth: 0.8052357876980987
AdaBoost: The AUC drops by: 0.018375621520431595
AdaBoost: AUC Score after excluding MentalHealth: 0.8201747530223082
AdaBoost: The AUC drops by: 0.003436656196222021
AdaBoost: AUC Score after excluding PhysicalHealth: 0.8225332734795201
AdaBoost: The AUC drops by: 0.0010781357390101665
AdaBoost: AUC Score after excluding HardToClimbStairs: 0.8208039767132937
AdaBoost: The AUC drops by: 0.002807432505236518
AdaBoost: AUC Score after excluding BiologicalSex: 0.8200444035058576
AdaBoost: The AUC drops by: 0.003567005712672655
AdaBoost: AUC Score after excluding AgeBracket: 0.8127518434526502
AdaBoost: The AUC drops by: 0.010859565765880097
AdaBoost: AUC Score after excluding EducationBracket: 0.8212673863206312
AdaBoost: The AUC drops by: 0.002344022897899034
AdaBoost: AUC Score after excluding IncomeBracket: 0.8201716445919043
AdaBoost: The AUC drops by: 0.003439764626625985
AdaBoost: AUC Score after excluding Zodiac: 0.8212848406687503
AdaBoost: The AUC drops by: 0.0023265685497799726
```

AUC Score for adaBoost : 0.8236114092185303

AdaBoost: AUC Score after excluding GeneralHealth: 0.8052357876980987

AdaBoost: The AUC drops by: 0.018375621520431595

Here we can see that the adaBoost improves the decision tree model even better than the random forest. The adaboost model can easily distinguish between classes.

Extra credit

#1: The best model of the 5 should be adaboost.

#2:The best predictor is generally GeneralHealth, but the condition of other predictors are quite strange. It seems that subject to different models, some predictors can either improve or worsen the performance of a model. For example, the predictor vegetable improves the SVM model performance but decreases the Decision Tree model performance.