Final Project : Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 69,973 Clinical Database Patient Records

Kanav Malik

April 29, 2020

Introduction

A final dataset of 69,973 observations and 10 dimensions is prepared for final analysis regarding the effect of HbA1C treatment on Readmission rates in diabetic patients.These consist of 9 potential explanatory variables namely gender,Race,Age,Diagnosis,Admissiom,Discharge, Medical_speciality,HbA1C,time_in_hospital and one response variable Readmitted

The report will compare models based on 3 approaches :

- Logistic regression model obtained following the inclusion instructions in the citation paper
- Logistic regression model made using my own model building strategy
- Random forest model

Test/Train Split(70/30 ratio)

The data is divided into 70% train data (48,106 observations) and 30% test data(21,867 observations)

The train data will be used for model building.The train data will be used for variable selection and calculating the parameter estimates of chosen variables

The test data will be used for model validation. This will be done by doing predictions for readmission on the test data. Then the validity of the model will be assessed by comparing the true values of readmission with the predicted values of readmissions to plot an area under curve and confusion matrix. Various statistics such as accuracy,sensitivity,specificity,Youden Index,Detection rate etc will be computed for each model. Youden Index will be used as the primary statistic to compare the models

Final Model as per citation paper

The following chunk of code was used to build the final model as per the citation paper:

```
model_final_paper <- glm(Readmitted ~ Discharge + Race + Admission + Medical_spec
  iality + time_in_hospital + Age + Diagnosis + HbA1C+ Discharge * Race + Dischar
  ge * Medical_speciality + Discharge * time_in_hospital + Discharge * Diagnosis
  + Medical_speciality * time_in_hospital + Medical_speciality * Age  + Medical_s
  peciality * Diagnosis + Admission * Age+  time_in_hospital * Diagnosis + Diagno
  sis * HbA1C, data = train, family = binomial)
```

Table 4 and 5(combined)

	Readmitted	
Predictors	*Log-Odds*	*p*
(Intercept)	-2.20	**<0.001**
DischargeOther	0.29	0.215
RaceCaucasian	0.02	0.760
RaceMissing	-0.45	**0.008**
RaceOther	-0.26	**0.031**
AdmissionOther	0.21	**0.022**
Admissionreferral	-0.02	0.785
Medical_specialityGeneral practice	-0.56	0.258
Medical_specialityInternal Medicine	-0.51	0.273
Medical_specialityMissing	-0.46	0.310

Medical_specialityOther	-1.00	**0.041**
Medical_specialitySurgery	-0.36	0.509
time_in_hospital	0.14	**<0.001**
Age[60,100)	0.12	0.491
Age<30	2.03	**0.032**
DiagnosisDigestive	-1.31	0.077
DiagnosisDisease of the Cirulatory	-0.94	**0.033**
DiagnosisGenitourinary	-0.98	0.219
DiagnosisInjury	-0.21	0.717
DiagnosisMusculoskeletal	-13.64	0.963
DiagnosisNeoplasms	-12.97	0.946
DiagnosisOther	-0.40	0.434
DiagnosisRespiratory	-1.07	**0.034**
HbA1CNormal result	-0.13	0.479
HbA1CResult High,changed	-0.61	**<0.001**
HbA1CResult High,not changed	-0.81	**0.003**
DischargeOther:RaceCaucasian	-0.02	0.855
DischargeOther:RaceMissing	0.37	0.111
DischargeOther:RaceOther	0.33	0.069
DischargeOther:Medical_specialityGeneral practice	0.24	0.275
DischargeOther:Medical_specialityInternal Medicine	0.15	0.466
DischargeOther:Medical_specialityMissing	0.16	0.387
DischargeOther:Medical_specialityOther	0.40	**0.045**
DischargeOther:Medical_specialitySurgery	0.54	**0.026**
DischargeOther:time_in_hospital	-0.04	**<0.001**
DischargeOther:DiagnosisDigestive	0.09	0.576
DischargeOther:DiagnosisDisease of the Cirulatory	0.12	0.368

DischargeOther:DiagnosisGenitourinary	-0.10	0.587
DischargeOther:DiagnosisInjury	0.47	**0.009**
DischargeOther:DiagnosisMusculoskeletal	0.56	**0.008**
DischargeOther:DiagnosisNeoplasms	-0.01	0.974
DischargeOther:DiagnosisOther	0.34	**0.015**
DischargeOther:DiagnosisRespiratory	0.33	**0.036**
Medical_specialityGeneral practice:time_in_hospital	-0.06	0.053
Medical_specialityInternal Medicine:time_in_hospital	-0.04	0.148
Medical_specialityMissing:time_in_hospital	-0.05	0.053
Medical_specialityOther:time_in_hospital	-0.05	0.055
Medical_specialitySurgery:time_in_hospital	-0.11	**0.004**
Medical_specialityGeneral practice:Age[60,100)	0.29	0.174
Medical_specialityInternal Medicine:Age[60,100)	0.29	0.135
Medical_specialityMissing:Age[60,100)	0.16	0.370
Medical_specialityOther:Age[60,100)	0.09	0.629
Medical_specialitySurgery:Age[60,100)	0.28	0.237
Medical_specialityGeneral practice:Age<30	-2.71	**0.015**
Medical_specialityInternal Medicine:Age<30	-2.06	**0.039**
Medical_specialityMissing:Age<30	-1.68	0.078
Medical_specialityOther:Age<30	-2.42	**0.012**
Medical_specialitySurgery:Age<30	-13.74	0.924
Medical_specialityGeneral practice:DiagnosisDigestive	1.22	0.113
Medical_specialityInternal Medicine:DiagnosisDigestive	1.12	0.133
Medical_specialityMissing:DiagnosisDigestive	1.08	0.144
Medical_specialityOther:DiagnosisDigestive	1.82	**0.017**
Medical_specialitySurgery:DiagnosisDigestive	0.89	0.269
Medical_specialityGeneral practice:DiagnosisDisease of the Cirulatory	1.14	**0.017**

Medical_specialityInternal Medicine:DiagnosisDisease of the Cirulatory	0.84	0.063
Medical_specialityMissing:DiagnosisDisease of the Cirulatory	0.84	0.055
Medical_specialityOther:DiagnosisDisease of the Cirulatory	1.22	**0.010**
Medical_specialitySurgery:DiagnosisDisease of the Cirulatory	0.39	0.450
Medical_specialityGeneral practice:DiagnosisGenitourinary	0.90	0.280
Medical_specialityInternal Medicine:DiagnosisGenitourinary	0.13	0.874
Medical_specialityMissing:DiagnosisGenitourinary	0.43	0.585
Medical_specialityOther:DiagnosisGenitourinary	1.11	0.171
Medical_specialitySurgery:DiagnosisGenitourinary	1.09	0.243
Medical_specialityGeneral practice:DiagnosisInjury	0.35	0.582
Medical_specialityInternal Medicine:DiagnosisInjury	-0.04	0.947
Medical_specialityMissing:DiagnosisInjury	-0.08	0.892
Medical_specialityOther:DiagnosisInjury	0.20	0.740
Medical_specialitySurgery:DiagnosisInjury	-0.32	0.616
Medical_specialityGeneral practice:DiagnosisMusculoskeletal	13.04	0.964
Medical_specialityInternal Medicine:DiagnosisMusculoskeletal	12.65	0.965
Medical_specialityMissing:DiagnosisMusculoskeletal	12.53	0.966
Medical_specialityOther:DiagnosisMusculoskeletal	13.13	0.964
Medical_specialitySurgery:DiagnosisMusculoskeletal	12.86	0.965
Medical_specialityGeneral practice:DiagnosisNeoplasms	12.92	0.946
Medical_specialityInternal Medicine:DiagnosisNeoplasms	13.38	0.944
Medical_specialityMissing:DiagnosisNeoplasms	13.20	0.945
Medical_specialityOther:DiagnosisNeoplasms	13.21	0.945
Medical_specialitySurgery:DiagnosisNeoplasms	12.00	0.950
Medical_specialityGeneral practice:DiagnosisOther	0.25	0.645
Medical_specialityInternal Medicine:DiagnosisOther	0.41	0.435
Medical_specialityMissing:DiagnosisOther	0.20	0.696

Medical_specialityOther:DiagnosisOther	0.84	0.116
Medical_specialitySurgery:DiagnosisOther	-0.04	0.942
Medical_specialityGeneral practice:DiagnosisRespiratory	0.35	0.527
Medical_specialityInternal Medicine:DiagnosisRespiratory	0.60	0.249
Medical_specialityMissing:DiagnosisRespiratory	0.54	0.286
Medical_specialityOther:DiagnosisRespiratory	1.13	**0.036**
Medical_specialitySurgery:DiagnosisRespiratory	0.82	0.221
AdmissionOther:Age[60,100)	-0.46	**<0.001**
Admissionreferral:Age[60,100)	-0.03	0.759
AdmissionOther:Age<30	0.08	0.836
Admissionreferral:Age<30	0.10	0.716
time_in_hospital:DiagnosisDigestive	-0.04	0.115
time_in_hospital:DiagnosisDisease of the Cirulatory	-0.03	0.092
time_in_hospital:DiagnosisGenitourinary	0.04	0.128
time_in_hospital:DiagnosisInjury	-0.04	0.097
time_in_hospital:DiagnosisMusculoskeletal	0.03	0.332
time_in_hospital:DiagnosisNeoplasms	-0.05	0.099
time_in_hospital:DiagnosisOther	-0.07	**0.002**
time_in_hospital:DiagnosisRespiratory	-0.03	0.254
DiagnosisDigestive:HbA1CNormal result	0.13	0.630
DiagnosisDisease of the Cirulatory:HbA1CNormal result	0.11	0.603
DiagnosisGenitourinary:HbA1CNormal result	0.49	0.095
DiagnosisInjury:HbA1CNormal result	-0.70	**0.038**
DiagnosisMusculoskeletal:HbA1CNormal result	0.16	0.625
DiagnosisNeoplasms:HbA1CNormal result	0.10	0.794
DiagnosisOther:HbA1CNormal result	0.02	0.932
DiagnosisRespiratory:HbA1CNormal result	-0.37	0.151

DiagnosisDigestive:HbA1CResult High,changed	0.97	**0.003**
DiagnosisDisease of the Cirulatory:HbA1CResult High,changed	0.81	**<0.001**
DiagnosisGenitourinary:HbA1CResult High,changed	0.40	0.340
DiagnosisInjury:HbA1CResult High,changed	0.21	0.637
DiagnosisMusculoskeletal:HbA1CResult High,changed	1.12	**0.007**
DiagnosisNeoplasms:HbA1CResult High,changed	0.15	0.787
DiagnosisOther:HbA1CResult High,changed	0.53	**0.033**
DiagnosisRespiratory:HbA1CResult High,changed	0.55	0.053
DiagnosisDigestive:HbA1CResult High,not changed	0.77	0.096
DiagnosisDisease of the Cirulatory:HbA1CResult High,not changed	0.67	**0.037**
DiagnosisGenitourinary:HbA1CResult High,not changed	0.32	0.634
DiagnosisInjury:HbA1CResult High,not changed	0.61	0.307
DiagnosisMusculoskeletal:HbA1CResult High,not changed	1.18	0.057
DiagnosisNeoplasms:HbA1CResult High,not changed	0.71	0.381
DiagnosisOther:HbA1CResult High,not changed	1.00	**0.004**
DiagnosisRespiratory:HbA1CResult High,not changed	0.45	0.299
Observations	48982	
Cox & Snell's R^2 / Nagelkerke's R^2	0.018 / 0.039	

Cut-off to be used for each class

Next we will check the cut-off values to be used based on the mean of each class

The following result is obtained :

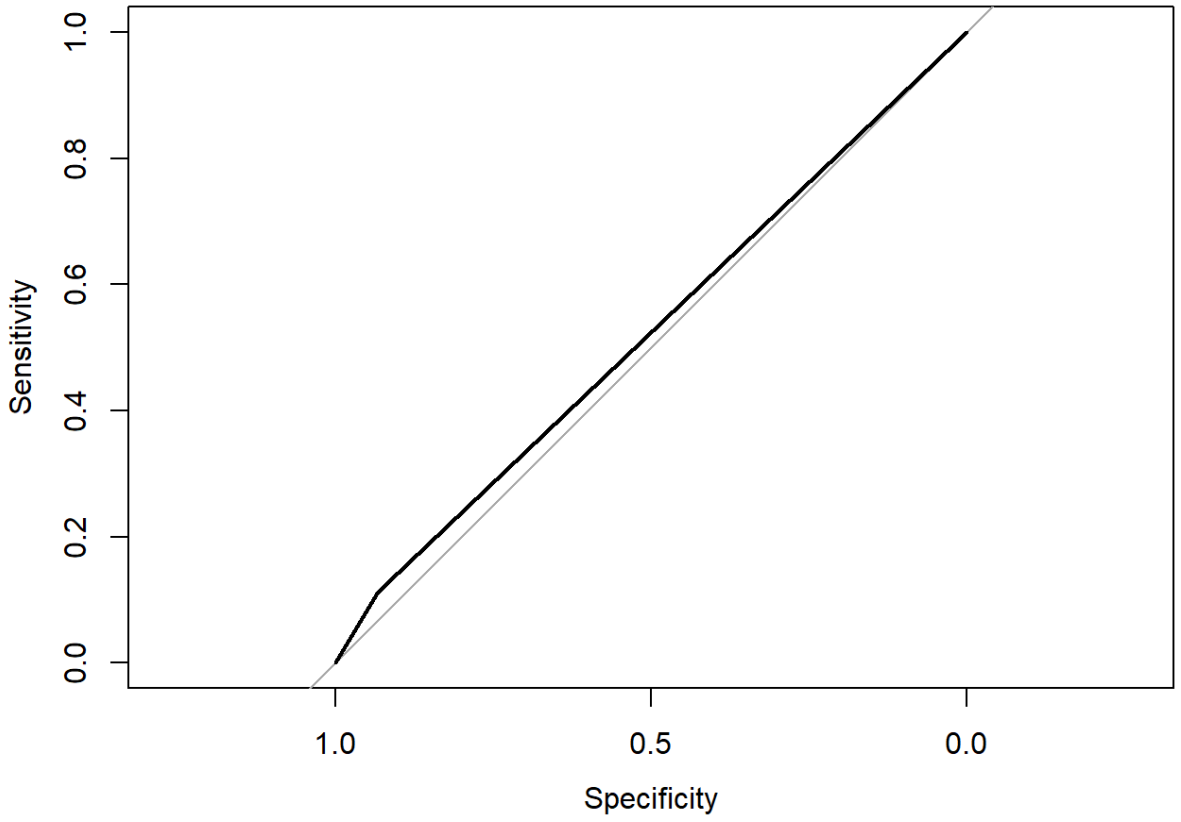
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.0607905 0.0834901 0.0905639 0.1173450 0.4355300
```

A cut-off value of 0.09 is used for class 1 and 0.91 is used for class 0 based on the predicted means above for each class

Model Validation

Plotting Area Under curve

```
## Area under the curve: 0.5225
```



Note : A cut-off value of 0.09 has been used above, as the proportion of readmissions in the train data is around 9%

The AUC is 52.25%

Confusion matrix

```
confusion_matrix <- table(predicted_readmission_rate$predicted,test$Readmitted)
confusionMatrix(confusion_matrix,positive = "1")
```

```
## Confusion Matrix and Statistics
##
##
##         0     1
##   0 10875   790
##   1  8275  1051
##
##                Accuracy : 0.5681
##                  95% CI : (0.5614, 0.5749)
##     No Information Rate : 0.9123
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0489
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.57089
##             Specificity : 0.56789
##          Pos Pred Value : 0.11270
##          Neg Pred Value : 0.93228
##              Prevalence : 0.08770
##          Detection Rate : 0.05007
##    Detection Prevalence : 0.44429
##       Balanced Accuracy : 0.56939
##
##        'Positive' Class : 1
##
```

The Accuracy is 56.81% whereas the Yuden Index is 0.134

The Sensitivity and Specificity are little imbalanced. The Youden Index is very low

The model has predicted False positives for 8,275 observations and False negatives for 790 observations

Out of the 20,991 observations in test data, 56.81% of the observations have been predicted correctly as being Readmitted or not

Specificity measures the proportion of actual negatives that are predicted as negative.The Specificity is 56.79%.This means that the model has true negative rate equal to 0.5679

Sensitivity measures the proportion of actual positives that are predicted as positive.The Sensitivity is 57.09%.This means that the model has true positive rate equal to 0.5709

Youden Index measures the combined effect of true positive rate(Sensitivity) and true negative rate(Specificity). As seen above, 0.134 on a scale of 0 to 1 is quite low

Overall, it is concluded that the odds of predicting readmission for this model is low

Building a regression model using my own model building strategy

A final model was tested using both AIC and BIC approach initially, The model made using AIC approach was quite similar to the model made in the citation paper. The model made using BIC approach outperformed the model made using AIC approach. The model using AIC approach has not been included in the report though it has been included in the R code provided separately. Below are the steps to arrive at the final logistic regression model using the BIC approach :

- Best subset selection using BIC criteria will be used to choose the explanatory variables which are important.HbA1C will be excluded from the best subset selection
- A model will be fitted including the selected explanatory variables from above step,HbA1C and their interactions. Type II chi-square test will be used to decide which interactions are important
- A final model will be made keeping all the significant variables and interactions

Best subset selection (BIC)

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Note: factors present with more than 2 levels.
```

After running the best subset selection using BIC criteria, the following predictors in the Table 4 are selected

Table 4

	y	
Predictors	*Log-Odds*	*p*
(Intercept)	-2.86	**<0.001**
Age[60,100)	0.23	**<0.001**
Age<30	-0.02	0.834
DischargeOther	0.54	**<0.001**
time_in_hospital	0.04	**<0.001**
Observations	48982	
Cox & Snell's R^2 / Nagelkerke's R^2	0.010 / 0.023	

Note: Log-Odds is the predictor parameter estimate

BIC criteria is a very parsimonious approach to variable selection. It imposes a higher penalty to a model with more variables

Based on the BIC criteria, the individual explanatory variables which are important are Age,Diagnosis and time_in_hospital

Next,we will check the significant interactions

Including the interactions

The following code was used:

```
model_interactions_BIC <- glm(Readmitted ~ Discharge + time_in_hospital + Age + H
  bA1C+ Discharge * time_in_hospital + Discharge * Age + time_in_hospital * Age +
  Discharge *HbA1C + time_in_hospital * HbA1C + Age * HbA1C , data = train, famil
  y = binomial)
```

The table below lists the parameter estimates and p-values for the model

Table 4 and 5 combined

	Readmitted	
Predictors	*Log-Odds*	*p*
(Intercept)	-2.97	**<0.001**
DischargeOther	0.75	**<0.001**
time_in_hospital	0.07	**<0.001**
Age[60,100)	0.25	**<0.001**
Age<30	0.19	0.378
HbA1CNormal result	-0.08	0.562
HbA1CResult High,changed	-0.13	0.417
HbA1CResult High,not changed	-0.02	0.937
DischargeOther:time_in_hospital	-0.04	**<0.001**
DischargeOther:Age[60,100)	-0.02	0.758
DischargeOther:Age<30	-0.38	0.260
time_in_hospital:Age[60,100)	-0.01	0.361
time_in_hospital:Age<30	0.02	0.708
DischargeOther:HbA1CNormal result	-0.03	0.812

DischargeOther:HbA1CResult High,changed	-0.09	0.537
DischargeOther:HbA1CResult High,not changed	0.17	0.420
time_in_hospital:HbA1CNormal result	-0.02	0.391
time_in_hospital:HbA1CResult High,changed	0.00	0.988
time_in_hospital:HbA1CResult High,not changed	-0.05	0.136
Age[60,100):HbA1CNormal result	0.12	0.371
Age<30:HbA1CNormal result	-0.41	0.399
Age[60,100):HbA1CResult High,changed	0.27	0.067
Age<30:HbA1CResult High,changed	-0.56	0.164
Age[60,100):HbA1CResult High,not changed	0.09	0.673
Age<30:HbA1CResult High,not changed	-0.67	0.133
Observations	48982	
Cox & Snell's R^2 / Nagelkerke's R^2	0.011 / 0.024	

Type II Chi-Square test is used to test the significance of each variable and interactions based on likelihood ratio test. It compares the ratio of deviance of model without the variable or interaction being tested to the deviance of the model including the variable or interaction being tested. A low p-value suggests that the variable or interactiob being tested has a significant effect on the model if it is removed. Therefore,it is important and should be kept in the model

Type II Chi Square Test

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Readmitted
##                            LR Chisq Df Pr(>Chisq)
## Discharge                   251.019  1  < 2.2e-16 ***
## time_in_hospital             46.742  1  8.099e-12 ***
## Age                          37.726  2  6.425e-09 ***
## HbA1C                         5.421  3  0.1434645
## Discharge:time_in_hospital   13.820  1  0.0002012 ***
## Discharge:Age                 1.360  2  0.5064998
## time_in_hospital:Age          1.140  2  0.5654907
## Discharge:HbA1C               1.131  3  0.7695641
## time_in_hospital:HbA1C        2.951  3  0.3992576
## Age:HbA1C                    10.647  6  0.0999301 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be observed from the Type II chi-square test that only one interaction namely Discharge * time_in_hospital is significant(p-value<0.01)

Final model using BIC approach initially

The below chunk of code is the final model which was made using the BIC criteria initially:

```
Model_Final_BIC <- glm(Readmitted ~ Discharge + time_in_hospital + Age + HbA1C +D
  ischarge * time_in_hospital, data = train, family = binomial)
```

The table below lists the parameter estimates and p-values for the final model made using the BIC criteria initially

Table 4 and 5 combined

	Readmitted	
Predictors	*Log-Odds*	*p*
(Intercept)	-2.94	**<0.001**
DischargeOther	0.73	**<0.001**
time_in_hospital	0.06	**<0.001**
Age[60,100)	0.22	**<0.001**
Age<30	0.00	0.999
HbA1CNormal result	-0.09	0.090
HbA1CResult High,changed	-0.02	0.733
HbA1CResult High,not changed	-0.17	0.098
DischargeOther:time_in_hospital	-0.04	**<0.001**
Observations	48982	
Cox & Snell's R^2 / Nagelkerke's R^2	0.011 / 0.024	

Finally a Type II Chi-Square test was performed to confirm whether the fitted variables and interaction are significant

Type II Chi Square Test

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Readmitted
##                            LR Chisq Df Pr(>Chisq)
## Discharge                   250.143  1  < 2.2e-16 ***
## time_in_hospital             47.332  1  5.994e-12 ***
## Age                          37.779  2  6.258e-09 ***
## HbA1C                         5.487  3     0.1394
## Discharge:time_in_hospital   15.925  1  6.590e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the fitted variables and interactions have a p-value<0.01 except HbA1C

We will keep HbA1C in our model as we need to see the impact of HbA1C on readmission rates, keeping other significant co-variates in the logistic regression model

Next, we will check how the fitted model performs on the test data

Cut-off to be used for each class

Next we will check the cut-off values to be used based on the mean of each class

The following result is obtained :

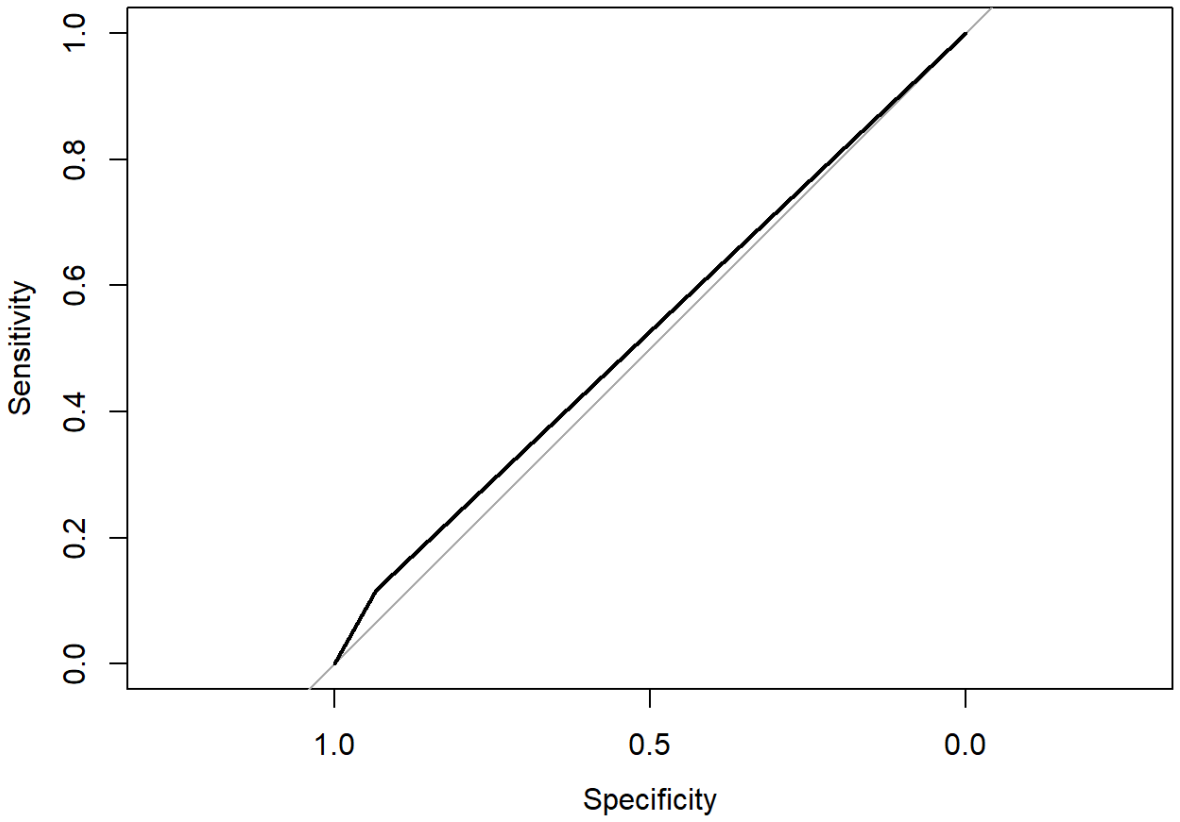
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04544 0.06568 0.07783 0.09056 0.12506 0.15196
```

A cut-off value of 0.09 is used for class 1 and 0.91 is used for class 0 based on the predicted means above for each class

Model Validation

Plotting Area Under curve

```
## Area under the curve: 0.5251
```



The Area under curve is 52.51% which is approximately equal to the final model as per the citation paper(52.16%)

Confusion matrix

```
## Confusion Matrix and Statistics
##
##
##         0     1
##   0 11436   819
##   1  7714  1022
##
##                Accuracy : 0.5935
##                  95% CI : (0.5868, 0.6001)
##     No Information Rate : 0.9123
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0566
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.55513
##             Specificity : 0.59718
##          Pos Pred Value : 0.11699
##          Neg Pred Value : 0.93317
##              Prevalence : 0.08770
##          Detection Rate : 0.04869
##    Detection Prevalence : 0.41618
##       Balanced Accuracy : 0.57616
##
##        'Positive' Class : 1
##
```

Note: For each statistic below, I will compare it to the statistic in the final model as per the citation paper

The Accuracy is 59.35%(vs 56.81%) whereas the Yuden Index is 0.152(vs 0.134)

The Sensitivity is 55.51%(vs 57.09%) which means that the model has a true positive rate of .5551(vs .5709) and Specificity is 59.72%(vs 56.79%) which means that the model has true negative rate equal to .5972 (vs 0.5679).The Yuden Index is still very low though it has improved slightly

The model has predicted False positives for 7,714 (vs 8,275) observations and False negatives in 819 (vs 790) observations

Out of the 20,991 observations in test data,59.35%(vs 56.81%) of the observations have been predicted correctly as being Readmitted or not

Overall, it is debatable whether this is sligthly improved model as compared to the model as per the citation paper.The odds of predicting readmission for this model is still low

Random Forest

A random forest classification model is made to predict the readmission rates.100 decision trees are made within the random forest. Same steps as above are followed for model validation of the random forest model

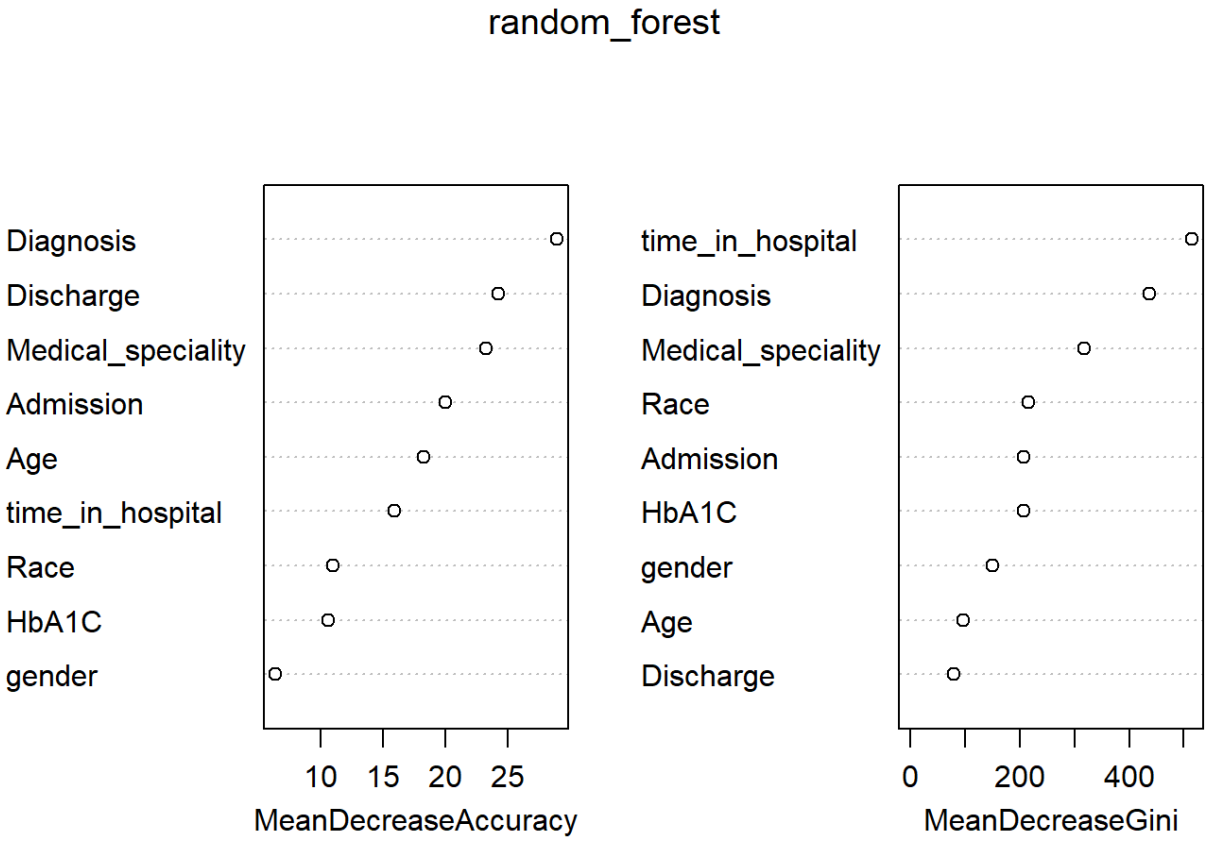
The below chunk of code was run to model a random forest with 100 trees:

```
random_forest <- randomForest(factor(Readmitted) ~ Discharge + Race + gender+ Adm
  ission + Medical_speciality + time_in_hospital + Age + Diagnosis + HbA1C, data
   = train, ntree = 100, importance = T)
```

A variable importance plot is made to access the importance of variables for the prediction of readmission through the random forest model above

Variable Importance Plot

```
##                    MeanDecreaseAccuracy
## Discharge                      24.228247
## Race                           10.995517
## gender                          6.381745
## Admission                      20.011172
## Medical_speciality             23.247224
## time_in_hospital               15.912090
## Age                            18.272276
## Diagnosis                      28.947094
## HbA1C                          10.604263
```



It can be seen from the variable importance plot above that removing gender,Race,HbA1C from the random forest model does not decrease the accuracy by much. However, a model without Race and gender make the predictions based on the random forest worse(model not included in the report). Therefore, they have been kept

Cut-off to be used for each class

Next we will check the cut-off values to be used based on the mean of each class

The following result is obtained :

```
##        0                1
##  Min.   :0.2703   Min.   :0.00000
##  1st Qu.:0.9706   1st Qu.:0.00000
##  Median :1.0000   Median :0.00000
##  Mean   :0.9713   Mean   :0.02870
##  3rd Qu.:1.0000   3rd Qu.:0.02941
##  Max.   :1.0000   Max.   :0.72973
```

A cut-off value of 0.03 is used for class 1 and 0.97 is used for class 0 based on the predicted means above for each class

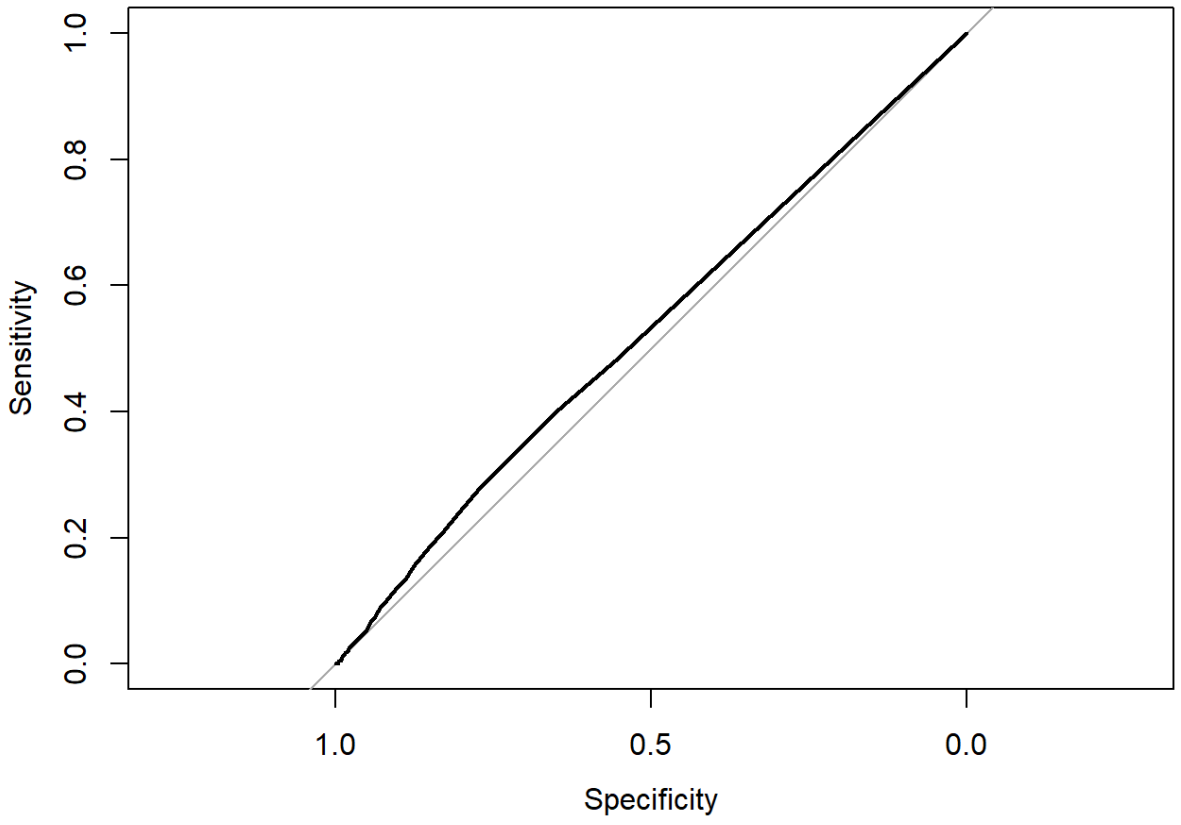
Random forest with cut-off

The following chunk of code was used to make the random forest with a cut-off of 0.03 for class 1 and 0.97 for class 0 :

```
random_forest_cutoff <- randomForest(factor(Readmitted) ~ Discharge + Race + gend
  er+ Admission + Medical_speciality + time_in_hospital + Age + Diagnosis + HbA1
  C, data = train, ntree = 100, importance = T,cutoff=c(0.97,0.03))
```

Model Validation

Plotting Area Under curve




```
## Area under the curve: 0.5265
```

The Area under curve is 52.65% which is slightly better than the AUC for the 2 models above

Confusion Matrix

```
##     0     1
## 15673  5318
```

```
## Confusion Matrix and Statistics
##
##
## predictions     0     1
##           0 14386  1287
##           1  4764   554
##
##                Accuracy : 0.7117
##                  95% CI : (0.7056, 0.7179)
##     No Information Rate : 0.9123
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0281
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.30092
##             Specificity : 0.75123
##          Pos Pred Value : 0.10417
##          Neg Pred Value : 0.91788
##              Prevalence : 0.08770
##          Detection Rate : 0.02639
##    Detection Prevalence : 0.25335
##       Balanced Accuracy : 0.52608
##
##        'Positive' Class : 1
##
```

The Accuracy has increased considerably as compared to the 2 models above. However the increase in accuracy has come at a huge cost

The sensitivity of the random forest is very low as compared to the 2 models above(31% vs around 56% for other two). Similarly, the Yuden Index has come down to 0.05 from 0.15 and 0.13

The model is better at predicting the True negatives as compared to the regression models above. However the model is poor at predicting the True positives. This is evident from the Detection rate too(which has fallen from around 5% for regression models to 2.6% here)

It can be concluded that although the accuracy and specificity of the model is high, it is poor at prediciting the positive value predicitions

Overall Conclusion

It is observed from the logisitic regression that the p-value for HbA1C is high. It should be fair to conclude that HbA1C does not have a significant impact on the readmission rates.This can be seen through the Type II Chi-Square tests that have been performed on the Logistic regression model made by me(p-value=0.1394)

The decision regarding which model is the best for predicting the readmission rates is debatable.It has been observed that the logisitic regression models are able to predict the True positives better,whereas the random forest is able to predict the True negatives better.Inherently, the class labels in the final dataset are imbalanced(class 0 has most of the observations).So the regression models and random forest have a lot of data for predicting the True negatives whereas the data is less for predicting the True positives.Maybe including more data for class label 1 or resampling the labels would help improve the model.

Another potential reason for the model being less efficient at prediciting the readmission rates might be that the explanatory variables included in the model are not sufficient to explain the variation in the readmission rates.A robust analysis of including more potential explantory variables in the final model and testing them for significance might help in improving the model.