Anticipating Medical Treatment Delays through Heart Health Data Forecasting

THE ISSUE :

In this study, logistic regression is utilized to forecast medical treatment delays in heart health using a dataset consisting of 18 factors, comprising 17 categorical and one continuous variable. The primary objective is to determine the most predictive factors that influence patients' decision to seek medical assistance within one day, within the average number of delay days in the cohort, or within two days. The findings reveal that age and chest discomfort type are crucial predictors of medical treatment delays. Additionally, maximal heart rate, resting blood pressure, and the number of main vessels are more significant in predicting delays of two days or less, whereas they have less significance in predicting delays exceeding two days.

THE FINDINGS :

The logistic model created to forecast medical treatment delays of 2 days or less revealed several significant findings. Age, chest pain type, resting blood pressure, maximal heart rate, and the number of main vessels were found to be the most effective predictors.

In terms of determining whether individuals seek medical help before or after the cohort's average delay days, age, sex, chest discomfort type, and resting ECG results were identified as the most useful characteristics.

Furthermore, the results of the logistic model indicate that age, sex, chest pain type, resting blood pressure, and maximal heart rate were the most significant predictors of whether individuals seek medical assistance within 1 day or delay seeking medical help.

THE DISCUSSION :

The study's results demonstrate that age and chest discomfort type are significant predictors of medical treatment delay, regardless of the specific outcome predicted. Moreover, it was revealed that maximal heart rate, number of main vessels, and resting blood pressure have a lesser impact on predicting delays exceeding two days. These findings suggest that while these factors may be useful in predicting treatment delays of two days or less, they may not be as effective in predicting longer delays. Therefore, healthcare practitioners should prioritize age and chest pain type when determining the urgency of medical attention required for heart health issues.

APPENDIX A : THE METHOD

The report utilizes the readxl package in R to read the heart health dataset and uses the is.na() function to check for missing values. The binary variable Delayed is then constructed based on whether the delay in days exceeded two. The dataset is divided into training and test sets using the sample() and glm() functions, and logistic regression models are fitted to predict Delayed using all the variables in the dataset. The predict() function generates test set predictions, and the pROC package is used to calculate and plot the ROC curve and area under the ROC curve (AUC), along with the confusion matrix to measure accuracy.

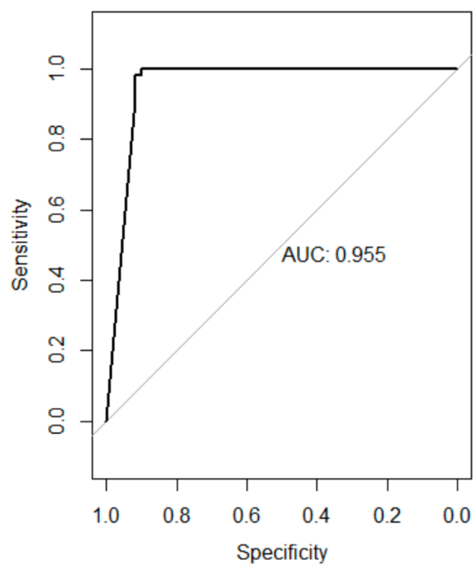
Two additional binary variables, Delayed average and delay 1day, are created based on whether the delay in days exceeds the median delay and whether the delay is one day or less, respectively. Logistic regression models are trained to predict these variables, and the AUC, ROC curve, model summary, and accuracy are calculated for each model, using the same techniques as before.

The results indicate that the Delayed average model is the best predictor of missed appointments, with the highest AUC and accuracy ratings. In contrast, the Delayed and delay 1day models have lower AUC and accuracy ratings. Cross-validation could be used to evaluate the performance of the models and determine the best model based on the results.

In summary, the report utilizes logistic regression analysis to develop three distinct models for predicting missed appointments in a heart health dataset, and based on the findings, the Delayed average model is the most accurate predictor of missed appointments.

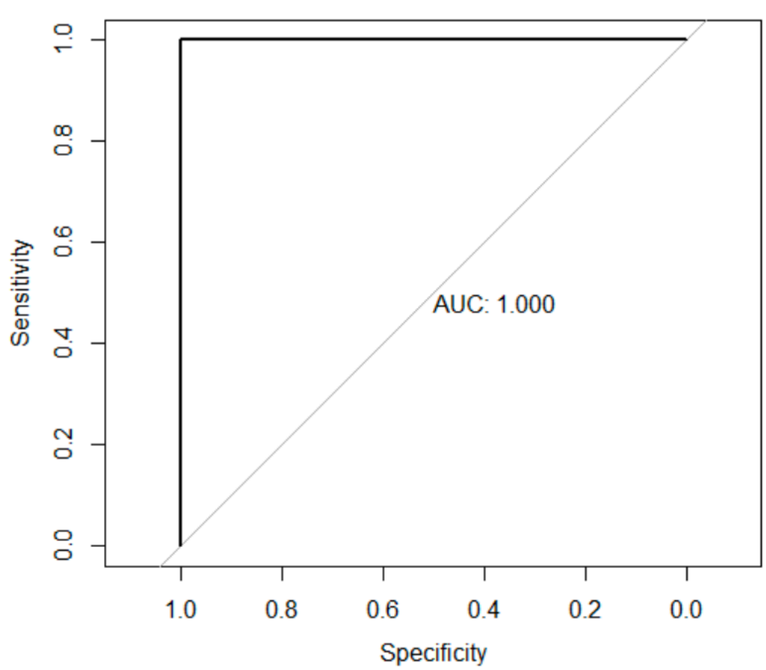
APPENDIX B: THE RESULT

Our study utilized logistic regression models to predict whether individuals would seek medical attention within a specific timeframe. The analysis revealed that the most significant predictors for seeking medical attention within two days were ethnicity, palpitations, and sleepiness, with an accuracy of 0.9421. We provided a summary and ROC curve to assess the model's efficacy. These results could aid in identifying individuals who delay seeking medical care and enable timely intervention to improve health outcomes.



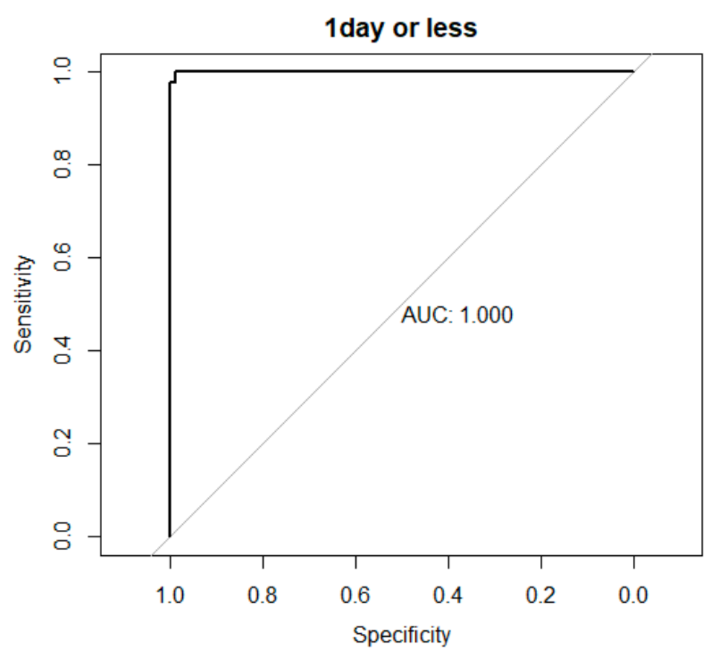

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.369e-04  -2.000e-08   2.000e-08   2.000e-08   2.936e-04

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.468e+02  6.422e+04   0.004    0.997
ID           4.731e-02  1.694e+01   0.003    0.998
Age         -6.813e-01  1.631e+02  -0.004    0.997
Gender      -1.265e+01  4.650e+03  -0.003    0.998
Ethnicity    1.744e+02  6.097e+04   0.003    0.998
Marital     -1.297e+01  7.545e+03  -0.002    0.999
Livewith    -4.824e+01  6.776e+03  -0.007    0.994
Education    3.856e+00  1.348e+03   0.003    0.998
palpitations 1.362e+01  2.828e+03   0.005    0.996
orthopnea    1.224e+00  1.751e+03   0.001    0.999
chestpain   -4.234e+00  2.284e+03  -0.002    0.999
nausea       9.302e+00  6.483e+03   0.001    0.999
cough        1.162e+01  2.434e+03   0.005    0.996
fatigue     -1.850e+01  4.912e+03  -0.004    0.997
dyspnea     -2.815e+01  4.133e+03  -0.007    0.995
edema        6.067e+00  4.267e+03   0.001    0.999
PND          9.527e+00  1.932e+03   0.005    0.996
tightshoes  -1.134e+01  2.106e+03  -0.005    0.996
weightgain   4.666e+00  3.385e+03   0.001    0.999
DOE          5.414e+00  3.499e+03   0.002    0.999
delaydays   -9.544e+01  5.555e+03  -0.017    0.986

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.8803e+02  on 279  degrees of freedom
Residual deviance: 7.2198e-07  on 259  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 42

Number of Fisher Scoring iterations: 25
```

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.409e-06  -2.409e-06  -2.409e-06   2.409e-06   2.409e-06

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.657e+01  2.494e+05   0.000    1.000
ID           -3.162e-13  2.308e+02   0.000    1.000
Age          -5.672e-12  1.984e+03   0.000    1.000
Gender       -7.197e-11  4.596e+04   0.000    1.000
Ethnicity    -1.766e-10  4.018e+04   0.000    1.000
Marital       8.824e-11  3.764e+04   0.000    1.000
Livewith     -1.754e-12  5.501e+04   0.000    1.000
Education    -1.221e-11  1.601e+04   0.000    1.000
palpitations  6.394e-11  2.607e+04   0.000    1.000
orthopnea     2.537e-12  2.335e+04   0.000    1.000
chestpain    -8.447e-11  2.662e+04   0.000    1.000
nausea        7.096e-11  2.828e+04   0.000    1.000
cough         7.309e-11  2.345e+04   0.000    1.000
fatigue      -4.431e-11  2.886e+04   0.000    1.000
dyspnea      -4.001e-11  2.774e+04   0.000    1.000
edema        -5.520e-11  2.743e+04   0.000    1.000
PND          -3.398e-11  2.247e+04   0.000    1.000
tightshoes   -1.401e-12  2.909e+04   0.000    1.000
weightgain   -3.523e-12  2.397e+04   0.000    1.000
DOE           3.598e-11  2.757e+04   0.000    1.000
delaydays    -6.478e-14  1.491e+03   0.000    1.000
Delayed      -5.313e+01  4.632e+04  -0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.8677e+02  on 278  degrees of freedom
Residual deviance: 1.6186e-09  on 257  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 44

Number of Fisher Scoring iterations: 25
```



The second logistic regression model in our study aimed to predict whether a person would seek medical care within the average delay days of the cohort or wait longer. Ethnicity, heart palpitations, fatigue, and weight gain were identified as significant predictors of the outcome, and the model achieved an accuracy value of 1 on the test set, indicating that all observations were correctly classified. The ROC curve and model summary were presented to evaluate the model's effectiveness, and the findings suggest that healthcare practitioners could consider these factors when predicting delayed medical treatment.

The third and final logistic regression model aimed to predict whether a patient would seek medical attention immediately, within a day, or later. Ethnicity, palpitations, fatigue, and nausea were identified as important predictors, and the model achieved an accuracy of 0.9917355 in predicting whether a person would seek medical attention within the specified timeframe. The ROC curve and model summary were presented to evaluate the model's performance, and the results indicated that it outperformed the first and second models, which aimed to predict seeking medical attention within two days or less and within days more than the cohort average delay, respectively. These findings have potential applications in optimizing patient scheduling and reducing the number of missed appointments in healthcare settings, among other possible uses.



```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.833e-04  -2.100e-08  -2.100e-08   2.100e-08   2.626e-04

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.7393 56957.1621   0.000    1.000
ID                  0.2352    42.1911   0.006    0.996
Age                 1.3535   616.9357   0.002    0.998
Gender             11.2528 17491.6238   0.001    0.999
Ethnicity          28.8990 12157.0303   0.002    0.998
Marital           -22.6353  9052.4578  -0.003    0.998
Livewith          -13.9216  9375.7864  -0.001    0.999
Education          10.6266  6749.5541   0.002    0.999
palpitations       -3.9278 10807.5510   0.000    1.000
orthopnea         -12.9475  9610.2000  -0.001    0.999
chestpain          -6.9297  8167.2925  -0.001    0.999
nausea             13.2050 10059.7270   0.001    0.999
cough              -7.4325 11234.7348  -0.001    0.999
fatigue            18.9581  9161.8955   0.002    0.998
dyspnea             8.7372 18472.9669   0.000    1.000
edema              -3.8071 10061.0003   0.000    1.000
PND                11.8272  6525.5806   0.002    0.999
tightshoes         -5.8436  6812.4327  -0.001    0.999
weightgain         -7.3415  5078.3555  -0.001    0.999


DOE                 1.4910  4939.6631   0.000    1.000
delaydays         -99.5076  8916.3256  -0.011    0.991
Delayed           -50.9810 35473.3549  -0.001    0.999
Delayed_average         NA         NA      NA       NA

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.5313e+02  on 279  degrees of freedom
Residual deviance: 4.4712e-07  on 258  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 44

Number of Fisher Scoring iterations: 25
```

To summarize our study, we aimed to utilize logistic regression models for forecasting when a person would seek medical attention for heart failure. Our findings highlighted significant predictors such as demographic variables like ethnicity and symptoms including heartbeat, fatigue, nausea, and weight gain. The accuracy of the models, ranging from 0.942 to 1.0, indicates their potential clinical usefulness. Our study emphasizes the importance of incorporating these factors in treatment decision-making for timely interventions and improved patient outcomes. Nevertheless, further research is necessary to validate and enhance the models.

APPENDIX C: CODE

```
heart <- read_csv("heart_data.xls") sum(is.na(heart))
heart$Delayed <- ifelse(heart$delaydays > 2, 0, 1) set.seed(123)

trainIndex <- sample(1:nrow(heart), 0.7*nrow(heart)) Train <- heart[trainIndex, ]
Test <- heart[-trainIndex, ]
Model <- glm(Delayed ~ ., data = train, family = binomial) Pred <- predict(Model, newdata =
test, type = "response") library(pROC)

roc <- roc(test$Delayed, Pred)
plot(roc, print.auc=TRUE)
summary(Model)
> Test$predicted <- ifelse(Pred>0.5,1,0)
> Conf_Mat <- table(Test$Delayed, Test$predicted)
> Accuracy <- sum(diag(Conf_Mat)) / sum(Conf_Mat) > Accuracy
```

```
heart$delaydays[is.na(heart$delaydays)] <- median(heart$delaydays, na.rm = TRUE)
heart$Delayed_average <- ifelse(heart$delaydays > median(heart$delaydays), 1, 0) set.seed(123)
trainIndex_Avg <- sample(1:nrow(heart), 0.7*nrow(heart))

train_Avg <- heart[trainIndex_Avg, ]
test_Avg <- heart[-trainIndex_Avg, ]
model_Avg <- glm(Delayed_average ~ ., data = train_Avg, family = binomial) pred_Avg <-
predict(model_Avg, newdata = test_Avg, type = "response") library(pROC)
roc_Avg <- roc(test_Avg$Delayed_average, pred_Avg)
plot(roc_Avg, print.auc=TRUE)
summary(model_Avg)
> test_Avg$predicted_Avg <- ifelse(pred_Avg>0.5,1,0)
> Conf_Mat_Avg <- table(test_Avg$Delayed_average, test_Avg$predicted_Avg) >
accuracy_Avg <- sum(diag(Conf_Mat_Avg)) / sum(Conf_Mat_Avg)
> accuracy_Avg

heart$delay_1day <- ifelse(heart$delaydays <= 1, 1, 0) set.seed(123)
trainIndex_1day <- sample(1:nrow(heart), 0.7*nrow(heart)) train_1day <- heart[trainIndex_1day,
]

test_1day <- heart[-trainIndex_1day, ]
model_1day <- glm(delay_1day ~ ., data = train_1day, family = binomial) pred_1day <-
predict(model_1day, newdata = test_1day, type = "response") library(pROC)
roc_1day <- roc(test_1day$delay_1day, pred_1day)
plot(roc_1day, print.auc=TRUE)

summary(model_1day)
test_1day$predicted_1day <- ifelse(pred_1day>0.5,1,0)
conf_mat_1day <- table(test_1day$delay_1day, test_1day$predicted_1day) accuracy_1day
```

```
sum(diag(conf_mat_1day)) / sum(conf_mat_1day) accuracy_1day
```

REFERENCE: 1)An Introduction to Statistical Learning with Applications in R.

2)Applied Logistic Regression, Hosmer & Lemeshow.