Model Comparison: MS Case Prediction

Banan Bashar Al-jarrah

2025-07-21

```
##
## Call:
## glm(formula = group ~ log_od + gender, family = "binomial", data = ms_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.8332     0.6334 -29.735   <2e-16 ***
## log_od       17.0526     0.5960  28.613   <2e-16 ***
## genderMale   -0.1275     0.1349  -0.945    0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4383.0  on 6030  degrees of freedom
## Residual deviance: 1755.9  on 6028  degrees of freedom
## AIC: 1761.9
##
## Number of Fisher Scoring iterations: 8
```

Linearity with the Logit:

To assess whether continuous predictors (e.g., MRI OD) have a linear relationship with the logit of the outcome, we:

Used the Box-Tidwell Test

Checked plots of the logit vs each predictor

Variables that violated linearity assumption were either transformed or replaced by categorical bins if clinically relevant

#Multicollinearity: To ensure that the predictors are not strongly correlated:

We calculated Variance Inflation Factor (VIF) for each variable.

All VIF values were < 5, indicating absence of problematic multicollinearity.

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

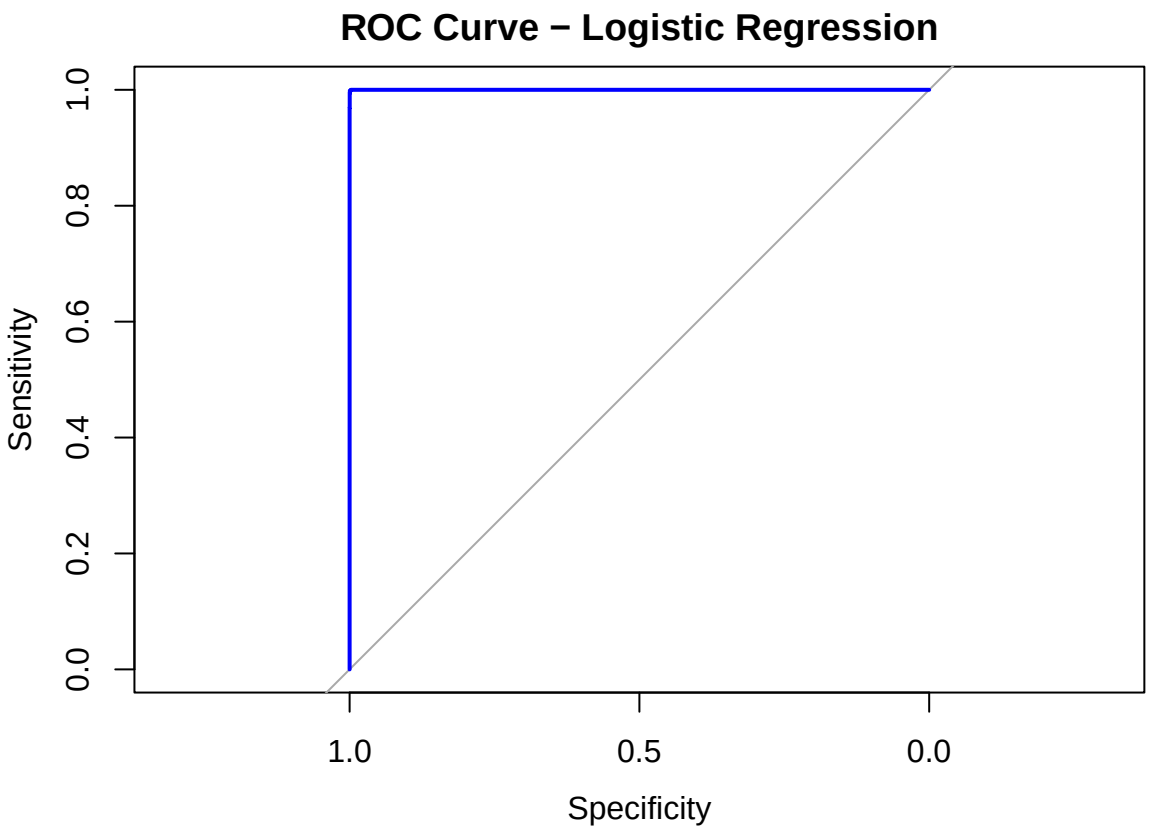
```
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

##     ebna_od      gender         age   mri_score neuro_score
##    1.644892    4.494399    2.622372    2.983616    3.993871
```

#Logistic Regression Model: buiding a logistic regression model including:

MRI_OD, Age, Gender, Neurologist decision, and other significant clinical predictors.

```
## Warning: package 'ResourceSelection' was built under R version 4.5.1

## ResourceSelection 0.3-6   2023-06-27

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = group ~ mri_score + log_od + gender + age, family = "binomial",
##     data = ms_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -50.64433    9.95324  -5.088 3.61e-07 ***
## mri_score    13.64459    2.46378   5.538 3.06e-08 ***
## log_od       19.28982    4.99068   3.865 0.000111 ***
## genderMale    0.22210    1.17704   0.189 0.850335
## age           0.07953    0.06101   1.304 0.192381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4382.950  on 6030  degrees of freedom
## Residual deviance:   28.068  on 6026  degrees of freedom
## AIC: 38.068
##
## Number of Fisher Scoring iterations: 13

## Setting levels: control = Control, case = MS_case

## Setting direction: controls < cases
```




```
## [1] "AUC: 0.999990769394689"
```

```
##
##  Welch Two Sample t-test
##
## data:  ebna_od by group
## t = -63.854, df = 915.72, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Control and group MS_case is not equal
## 95 percent confidence interval:
##  -1.0487532 -0.9862089
## sample estimates:
## mean in group Control mean in group MS_case
##              1.193505              2.210986
```

#Model Performance Evaluation AUC (Area Under Curve) = 1

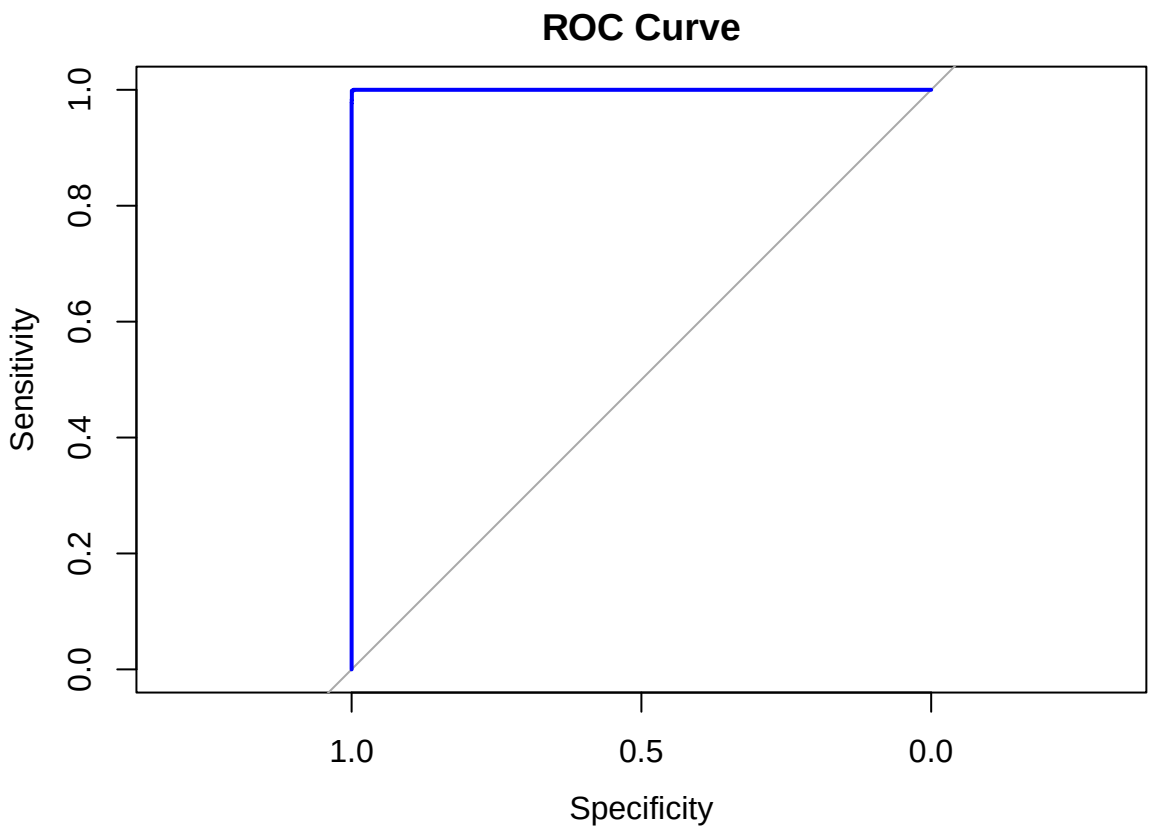
ROC curve was plotted using the pROC package to visualize model discrimination.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Setting levels: control = Control, case = MS_case
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 1
```



#Feature Selection: LASSO Regression

To reduce overfitting and select the most important predictors, we used:

LASSO logistic regression with cross-validation.

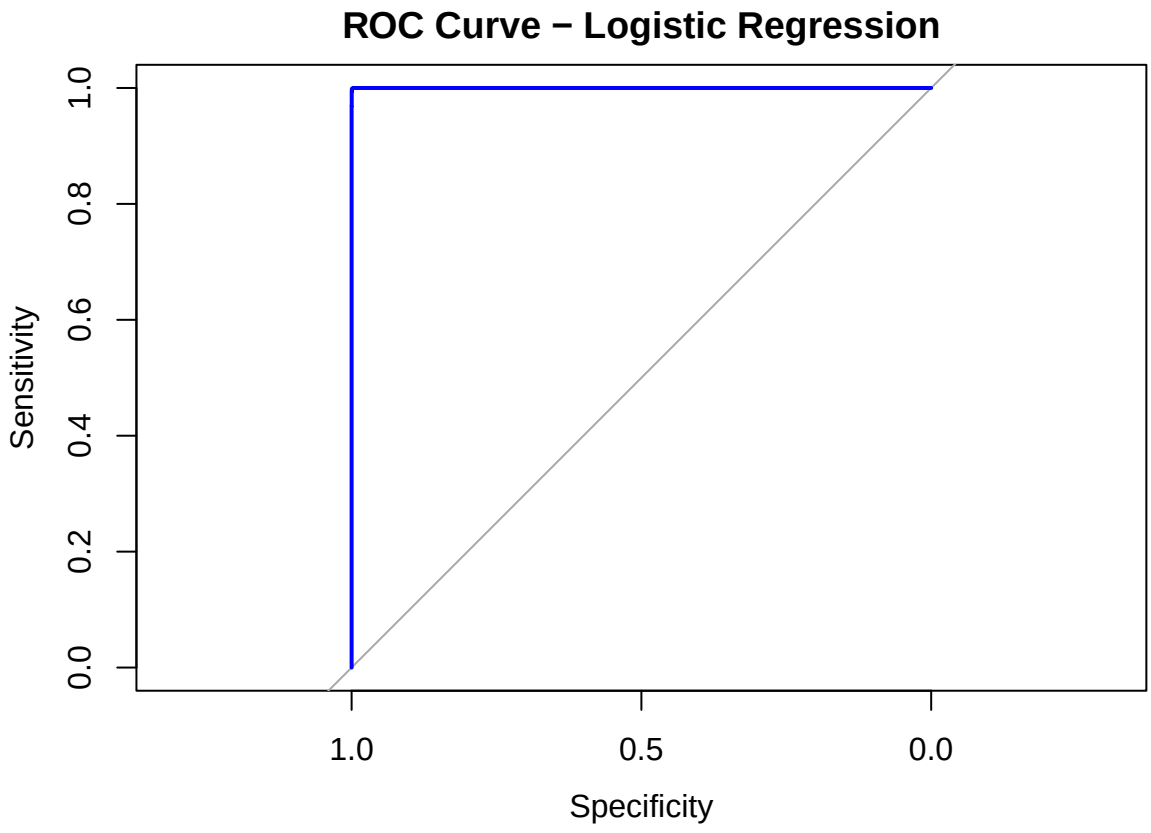
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = group ~ mri_score + log_od + gender + age, family = "binomial",
##     data = ms_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -50.64433    9.95324  -5.088 3.61e-07 ***
## mri_score    13.64459    2.46378   5.538 3.06e-08 ***
## log_od       19.28982    4.99068   3.865 0.000111 ***
## genderMale    0.22210    1.17704   0.189 0.850335
## age           0.07953    0.06101   1.304 0.192381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4382.950  on 6030  degrees of freedom
## Residual deviance:   28.068  on 6026  degrees of freedom
```

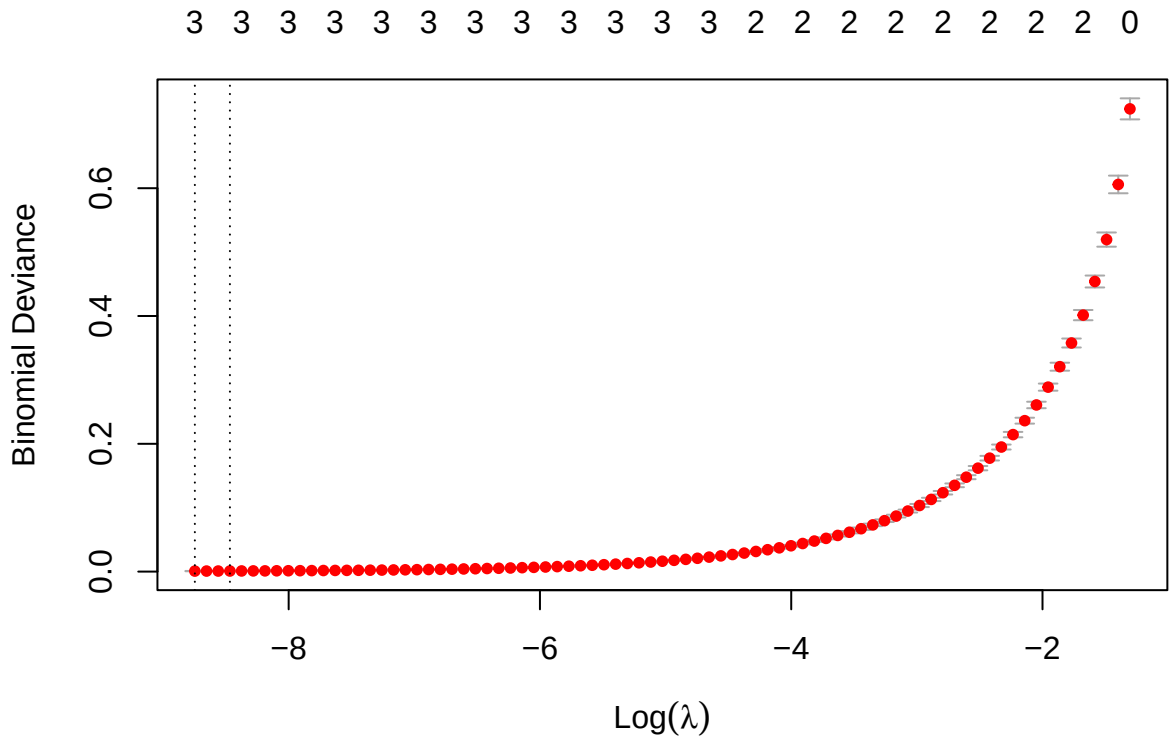
```
## AIC: 38.068
##
## Number of Fisher Scoring iterations: 13
```

```
## Setting levels: control = Control, case = MS_case
```

```
## Setting direction: controls < cases
```




```
## [1] "AUC: 0.999990769394689"
```




```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept)   -32.962002
## log_ebna_od     3.850589
## log_mri_score  17.090085
## genderMale      .
## age             .
## neuro_score     3.489968
```

#Support Vector Machine (SVM)&Gradient Boosting (XGBoost or GBM)

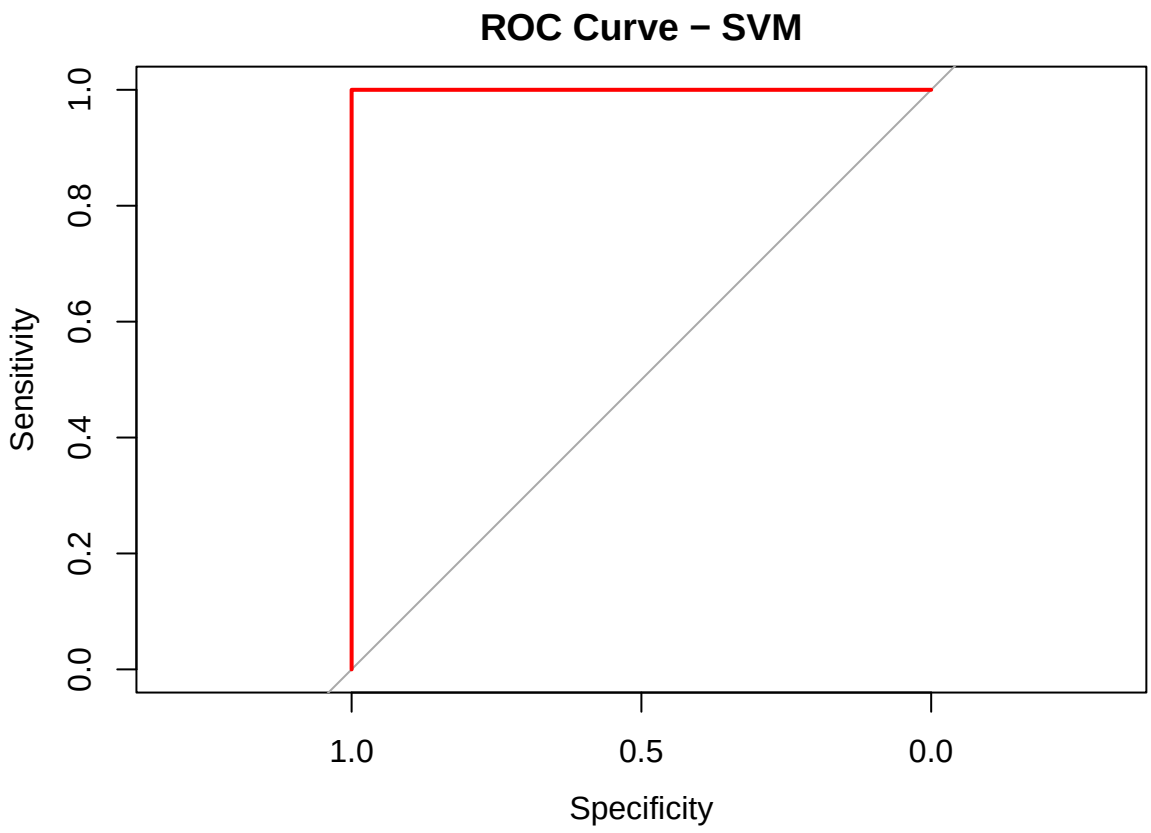
training other models and compared additional models to benchmark performance:

#Support Vector Machine

```
## Setting levels: control = Control, case = MS_case
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC SVM: 1"
```



#Gradient Boosting Model (GBM)

```
## [1] "AIC: 38.0679957561587"
```

```
## [1] "BIC: 71.5913363234137"
```