Homework 5

Rylie Fleckenstein

2/17/21

1) Question 4.7.6 pg 170 show your work

Suppose we collect data for a group of students in a statistics class with variables X_1=hours, X_2=undergrad GPA, and Y=receive an A. We fit a logistic regression and produce an estimated coefficient $\beta_0 = -6, \beta_1 = 0.05, \beta_2 = 1$

The equation for used in multiple logistic regression for predicting the probability of the binary response is the following:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

for p predictors.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

We know from the question that $X_1 = 40, X_2 = 3.5$. If we take these parameters and input them into the multiple logistic regression function we get the following:

$$p(X) = \frac{e^{-6+0.05(40)+1(3.5)}}{1+e^{-6+0.05(40)+1(3.5)}} = \frac{e^{-0.5}}{1+e^{-0.5}}$$

$$p(X) = 0.37754$$

Therefore, there is a 37.8% chance the a student who studied for 40 hours and had an undergrad GPA of 3.5 would receive an A for the class in question.

(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

In order to determine how many hours the student from part A would need to study in order to have a 50% chance of receiving an A we solve for x in the following equation:

$$0.5 = \frac{e^{-6+0.05(x)+1(3.5)}}{1 + e^{-6+0.05(x)+1(3.5)}}$$

$$0.5(1 + e^{0.05(x)-2.5}) = e^{0.05(x)-2.5}$$

$$0.5 + 0.5(e^{0.05(x)-2.5}) = e^{0.05(x)-2.5}$$

$$0.5 = 0.5(e^{0.05(x)-2.5})$$

$$1 = e^{0.05(x)-2.5}$$

$$ln(1) = ln(e^{0.05(x)-2.5})$$

$$.05x - 2.5 = 0$$

$$x = 50$$

So, in order for this student to have a 50% chance of receiving an A, they need to study for 50 hours.

2) Continue from Homework 3 and 4 using the Weekly data set from 4.7.10). Construct a model (using the predictors chosen for previous homework) and fit this model using the MclustDA function from the mclust library.

i) Provide a summary of your model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n     % Model G
##    Down 441 44.77     V 2
##    Up   544 55.23     V 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
## Classification error = 0.4416
## Brier score          = 0.2452
```

- What is the best model selected by BIC? Report the Model Name and the BIC. (See mclustModelNames)

The model chosen by the MclustDA function for this particular problem was the “V” model. According to the documentation that stands for a “variable variance (one-dimensional)” model. The BIC score for this model was -4327.804.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

From above model summary we can see that the training error for this model is at 44.16%. Below I have utilized a function that I wrote to calculate and print the test error, true positive rate (sensitivity), and the true negative rate (specificity). We can see that the test error is at 45.19%, the true positive rate is at 11.62%, the true negative rate is at 85.24%, and then the overall correct classification rate is at 54.8%. Again, the mclustDA model continues to follow the same trend as the models in homework 3 and 4 by being able to predict when the market goes up but struggling to make accurate predictions for when the market will go down.

```
## $‘Confusion Matrix‘
##
##        Down Up
##   Down    5 38
##   Up      9 52
##
## $‘Overall Percentage Correct‘
## [1] 54.80769
##
## $Sensitivity
## [1] 11.62791
##
## $Specificity
## [1] 85.2459
##
## $‘Test Error‘
## [1] 45.19231
```

ii) Repeat the `MclustDA` analysis, but this time specify `modelType = "EDDA"`. Provide a summary of this model.

Here we are repeating the MclustDA analysis but are making one change to the model structure. We are setting modelType = “EDDA”. What this does is it constrains the model so that it only has a single component for each class with the same covariance structure among classes.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
```

```
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -2204.237 985  3 -4429.152
##
## Classes   n     % Model G
##    Down 441 44.77     E 1
##    Up   544 55.23     E 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   22 419
##   Up     20 524
## Classification error = 0.4457
## Brier score          = 0.2462
```

- What is the best model using BIC as the model selection criteria?

The model chosen for use in this mixture model is the "E" model. According to the documentation that stands for "equal variance (one-dimensional)" model.

The BIC for this MclustDA model comes in at -4429.152 which is lower than the first MclustDA model we fit (-4327.804). This suggests that if we were to use BIC as the model selection criteria, this model if better than the first MclustDA model.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

As we see below, the training error for this model was 44.57%, the test error is coming in at 37.5%, true positive 20.9%, true negative 91.8%, and then overall correct classification is at 62.5%.

```
## $`Confusion Matrix`
##
##        Down Up
##   Down    9 34
##   Up      5 56
##
## $`Overall Percentage Correct`
## [1] 62.5
##
## $Sensitivity
## [1] 20.93023
##
## $Specificity
## [1] 91.80328
##
## $`Test Error`
## [1] 37.5
```

iii) Compare the results with Homework #3 & 4. Which method performed the best? Justify your answer. *Present your results in a well formatted table; include the previous methods and their corresponding rates.*

If we take a look at the table below we can see that we have 3 different models that have performed identically. The three models in question are the logistic regression model from homework 3, the LDA model from homework 4, and the MclustDA model from homework 5. Since they have all performed relatively the same, it is potentially the best practice to pick the simplest model of the three. Using that logic, I would have to chose the logistic regression model as the final model.

Table 1: Model Scores

Models	Accuracy	Sensitivity	Specificity	Test.Error
Log Reg	62.50	20.93	91.80	37.50
LDA	62.50	20.93	91.80	37.50
QDA	58.65	0.00	100.00	41.34
KNN	59.61	51.28	64.61	40.38
MclustDA	62.50	20.93	91.80	37.50

iv) From the original model variables, construct a new set of variables, fit a model using `MclustDA` and repeat i-iii. *Hint: new variables may be interactions, polynomials, and/or splines.* Do these new variables give an improvement in error rates compared to previous models? Explain how the new variables were constructed.

I constructed a matrix of transformed variables to be used in the MclustDA function. I played around with different transformations of the variables to include taking the log() or creating polynomials of different variables. I eventually found a combination that seems to work relatively well which utilizes some variable interactions. I found that by creating an interaction between Lag1 * Lag2, and another interaction between Lag2 * Lag3 we are able to create a model that performs almost as well as the previous best models. In order to construct these variables I utilized the cbind() function in order to create a matrix of new variables. I then created an interaction between the terms by using the "*" function. I did this for the training set and the test set so that both were transformed in the same way. The final model with the transformed variables was able to achieve an accuracy of 61.54% which is very close to the best models of before that had an accuracy of 62.5%. The major differences between this model and those models shows itself in the true positive and true negative rates. This model with the transformed data was able to out predict the previous models when it came to specificity (true negative rate). It achieved a rate of 93.44% where the previous models only achieved 91.8%. This model, however, did falter on its sensitivity rate (true positive) by only achieving 16.27% where as the previous achieved 20.93%. This type of improvement can be useful in some situations where its not a big deal to underestimate a class in return for more accurately predicting the other.

i) Provide a summary of your model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df      BIC
##       -4985.125 985 34 -10204.6
##
## Classes   n     % Model G
##    Down 441 44.77   VII 4
##    Up   544 55.23   VII 5
##
```

```
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  151 290
##   Up    147 397
## Classification error = 0.4437
## Brier score          = 0.2463
```

```
## $‘Confusion Matrix‘
##
##        Down Up
##   Down   11 32
##   Up     17 44
##
## $‘Overall Percentage Correct‘
## [1] 52.88462
##
## $Sensitivity
## [1] 25.5814
##
## $Specificity
## [1] 72.13115
##
## $‘Test Error‘
## [1] 47.11538
```

- What is the best model selected by BIC? Report the Model Name and the BIC. (See mclustModelNames)

The best model selected by the BIC was the “VII” model which, according to the documentation, stands for spherical, unequal volume. The BIC for this model came out to be -10204.6.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

From calling the summary of the model and the function that I wrote we can see that the training error for this model was at 44.37%, the test error was at 47.11%, true positive 25.58%, and true negative at 72.13%.

ii) Repeat the `MclustDA` analysis, but this time specify `modelType = "EDDA"`. Provide a summary of this model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -7175.326 985  8 -14405.79
##
## Classes   n     % Model G
##    Down 441 44.77   EVE 1
##    Up   544 55.23   EVE 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   15 426
##   Up     31 513
## Classification error = 0.464
## Brier score          = 0.2543
```

```
## $‘Confusion Matrix‘
##
##        Down Up
##   Down    7 36
##   Up      4 57
##
## $‘Overall Percentage Correct‘
## [1] 61.53846
##
## $Sensitivity
## [1] 16.27907
##
## $Specificity
## [1] 93.44262
##
## $‘Test Error‘
## [1] 38.46154
```

- What is the best model using BIC as the model selection criteria?

The best model using BIC was the “EVE” which stands for ellipsoidal, equal volume and orientation (*). The BIC for this model was at -14405.79 which beats the previous model. Therefor, based upon the BIC, this model should perform better then the first one with the transformed variables.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

According to the summary and the function I wrote the training error for this model is coming in at 46.4%, the test error is 38.46%, the true positive is 16.27%, and the true negative is 93.44%.

iii) Compare the results with Homework #3 & 4. Which method performed the best? Justify your answer. *Present your results in a well formatted table; include the previous methods and their corresponding rates.*

The best model still stands as being the logistic regression model for the simple reason that it is the simplest model and did score the best. There are some situations where we might want to use the final MClustDA model which has a higher specificity rate but generally speaking logistic regression is the clear winner of models for this group.

Table 2: Model Scores

Models	Accuracy	Sensitivity	Specificity	Test.Error
Log Reg	62.50	20.93	91.80	37.50
LDA	62.50	20.93	91.80	37.50
QDA	58.65	0.00	100.00	41.34
KNN	59.61	51.28	64.61	40.38
MclustDA	61.54	16.27	93.44	38.46

Sources:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer