**STAT 206**

**Take-home Final**

**Please submit a Microsoft word copy of your answer to BBlearn by August 29th before 11:59 PM. Late submission is not accepted.**

1. **Predicting Wine Quality (10 pts each)**

Descriptions:

This dataset is related to red vinho verder wine samples from the north of Portugal. The goal is to model wine quality(v12) based on physicochemical attributes (v1-v11). Each row of the data set corresponds to a wine sample. Descriptions for the data follow:

|  |  |
| --- | --- |
| v1 | fixed acidity |
| v2 | volatile acidity |
| v3 | critic acid |
| v4 | residual sugar |
| v5 | chlorides |
| v6 | free sulfur dioxide |
| v7 | total sulfur dioxide |
| v8 | density |
| v9 | pH |
| v10 | sulphates |
| v11 | alcohol |
| v12 | quality (score from 1-10) |

Please download the **train.csv** and **test.csv** from BBlearn. A wine sample is identified as good wine if the quality v12>=6. The research goal is, therefore, to predict if a wine sample is a good wine. Please create a new dependent variable that takes the values of 1 if a sample is good wine and 0 otherwise.

1. Please build a **logistic regression** model that includes all the attributes from the training set to predict the wine quality (model 1) in the testing set. Show the confusion table. What are precision, recall, and accuracy?

By training the logistic regression model on the training set and test on test set, the below confusion table will be obtained:

|  |  |  |
| --- | --- | --- |
| Predict/Actual | Not Good Wine | Good Wine |
| Not Good Wine | 54 | 17 |
| Good Wine | 22 | 67 |

From the above confusion table, the precision rate = TP/(TP+FP) = 67/(67+22) = 75.28%,

Recall rate = TP/(TP+FN) = 67/(67+17) = 79.76%, Accuracy rate = (54+67)/(54+67+22+17) = 121/160 = 75.63%.

1. From model 1, what will happen to the odds of good wine if the concentration of sulphates is increased by 1 unit?

Here is the summary of model 1:

Call:

glm(formula = wine ~ ., family = binomial, data = train2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4792 -0.8348 0.3169 0.8312 2.3544

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 31.701830 83.879004 0.378 0.7055

v1 0.118951 0.103948 1.144 0.2525

v2 -3.548543 0.525526 -6.752 1.45e-11 \*\*\*

v3 -1.427898 0.599317 -2.383 0.0172 \*

v4 0.058300 0.055385 1.053 0.2925

v5 -3.763745 1.654915 -2.274 0.0229 \*

v6 0.020680 0.008769 2.358 0.0184 \*

v7 -0.016377 0.003108 -5.269 1.37e-07 \*\*\*

v8 -39.243720 85.646802 -0.458 0.6468

v9 -0.447933 0.755797 -0.593 0.5534

v10 2.705454 0.477189 5.670 1.43e-08 \*\*\*

v11 0.887090 0.109353 8.112 4.97e-16 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

From the above summary of model 1, it can be observed that if the concentration of sulphates(v10) is increased by 1 unit, the odds of good wine will be increased by e^2.705 = 14.95 times to the odds of bad wine.

1. Please build a **logistic regression** model that v1, v2, v3, v4, and v5 from the training set to predict the wine quality (model 2) in the testing set. Show the confusion table. What are precision, recall, and accuracy?

By applying the model 2, we can get the following confusion table:

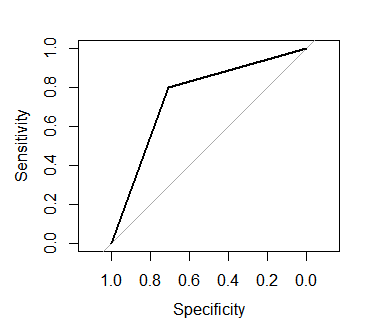
|  |  |  |
| --- | --- | --- |
| Predict/Actual | Not Good Wine | Good Wine |
| Not Good Wine | 46 | 27 |
| Good Wine | 30 | 57 |

From the above confusion table, the precision rate = TP/(TP+FP) = 57/(57+30) = 65.52%,

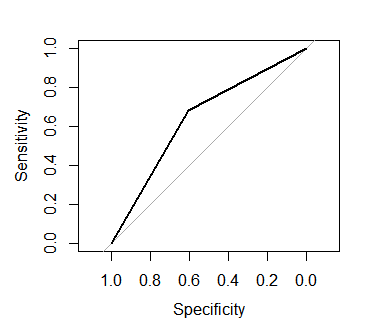
Recall rate = TP/(TP+FN) = 57/(57+27) = 67.86%, Accuracy rate = (46+57)/(46+57+27+30) = 121/160 = 64.38%.

1. Please compare model 1 and model 2 using the ROC curve on the testing set. Which one performs better?

ROC Curve for Model 1:



ROC Curve for Model 2:



By comparing the two ROC curves for model 1 and model 2, it is obvious that model 1 has a better performance than model 2.

1. Suppose that a merchant will lose $10 if he misclassifies a bad wine as a good wine. He will also lose $1 if he misclassifies a good wine as a bad wine. Please compare the total cost of model 1 and model 2.

By using the confusion table, we can get:

For model 1: cost = 10\*FP + 1\*FN = 10\*22 + 1 \* 17= 237

For model 2: cost = 10\*FP + 1\*FN = 10\*30 + 1\* 27 = 327

It can be compared that the cost for model 1 is much lower than that for model 2.

1. We can also use a linear regression model for classification. Please use v12 as the dependent variable and build a **linear regression model** on all the attributes from the training set. Please show the Mean Square Error (**MSE**): the average of the squared differences between the predicted and actual wine quality score.

Below is the result of linear regression model on all attributes for training set:

Call:

lm(formula = v12 ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-2.66387 -0.36046 -0.05331 0.44620 2.02054

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19.673870 22.267169 0.884 0.3771

v1 0.025452 0.027519 0.925 0.3552

v2 -1.116973 0.127878 -8.735 < 2e-16 \*\*\*

v3 -0.227165 0.155928 -1.457 0.1454

v4 0.016684 0.015472 1.078 0.2811

v5 -1.887863 0.436970 -4.320 1.67e-05 \*\*\*

v6 0.003400 0.002289 1.485 0.1377

v7 -0.003130 0.000777 -4.028 5.92e-05 \*\*\*

v8 -15.550769 22.731083 -0.684 0.4940

v9 -0.422898 0.199436 -2.120 0.0341 \*

v10 0.888312 0.119437 7.437 1.76e-13 \*\*\*

v11 0.281376 0.027409 10.266 < 2e-16 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.6451 on 1427 degrees of freedom

Multiple R-squared: 0.3619, Adjusted R-squared: 0.357

F-statistic: 73.57 on 11 and 1427 DF, p-value: < 2.2e-16

Thus the MSE = Average(sum of squared residuals) = 0.4127

1. We can also use a linear regression model for classification. Please use v12 as the dependent variable and build a **linear regression model** on all the attributes from the training set. To make decisions, a merchant will classify a sample as good wine if the predicted V12 is greater than or equal to 6. Show the confusion table.

Confusion table based on Linear regression model:

|  |  |  |
| --- | --- | --- |
| Predict/Actual | Not Good Wine | Good Wine |
| Not Good Wine | 72 | 50 |
| Good Wine | 4 | 34 |

1. It’s very important to inspect the dataset before conducting analysis. Please download the **final.csv** from BBlearn. Use v10 as the dependent variable and build a **logistic regression model** on all the attributes. You may find that the algorithm does not converge as the warning message indicates. Can you find out why? (**Extra 1%**)

> final.logit = glm(v10~., data=final, family = binomial)Warning message:glm.fit: algorithm did not converge > summary(final.logit)

Call:

glm(formula = v10 ~ ., family = binomial, data = final)

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 9.465e+06 0 1

v1 2.289e-08 8.760e+04 0 1

v2 -6.459e-13 1.006e+04 0 1

v3 -3.091e-07 3.531e+05 0 1

v4 5.834e-10 1.664e+03 0 1

v5 -3.927e-10 6.172e+02 0 1

v6 3.331e-06 9.402e+06 0 1

v7 3.837e-08 1.046e+05 0 1

v8 -8.102e-08 8.496e+04 0 1

v9 2.387e-09 1.608e+04 0 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 0.0000e+00 on 854 degrees of freedom

Residual deviance: 4.9604e-09 on 845 degrees of freedom

AIC: 20

Number of Fisher Scoring iterations: 25

Gives the warning message that the algorithm did not converge:

The reason why appear this message id that the algorithm hit the maximum number of allowed iterations(default) before coming to convergence. We can set higher iteration number to resolve this problem.