In the scorecard development, the area under ROC curve, also known as AUC, has been widely used to measure the performance of a risk scorecard. Given everything else equal, the scorecard with a higher AUC is considered more predictive than the one with a lower AUC. However, little attention has been paid to the statistical analysis of AUC itself during the scorecard development.

While it might be less of a concern to rely on a simple comparison of AUC for the model selection in the development stage and then to pick the scorecard with a higher AUC, more attention should be called for on AUC analysis in the post-development stage. For instance, the senior management would need to decide whether it is worthy to retire a legacy scorecard that might be still performing and to launch the full-scale deployment of a new scorecard just for an increase in AUC that might not even be statistically significant. While the claim of certain business benefits can always be used as an argument in favor of the new scorecard, the justification would become even more compelling with a solid statistical evidence. What’s more, the model validation analyst might also want to leverage the outcome of AUC analysis to ensure the statistical soundness of new scorecards.

In the example below, two logistic regressions were estimated with AUC = 0.6554 and BIC = 6,402 for the model with 6 variables and AUC = 0.6429 and BIC = 6,421 for the model with 3 variables.

df1 <- read.csv("Downloads/credit\_count.txt")

df2 <- df1[which(df1$CARDHLDR == 1), ]

y <- "DEFAULT"

x1 <- c("OWNRENT", "INCOME", "INCPER", "LOGSPEND", "AGE", "EXP\_INC")

x2 <- c("MAJORDRG", "MINORDRG", "INCOME")

m1 <- glm(eval(paste(y, paste(x1, collapse = " + "), sep = " ~ ")), data = df2, family = binomial)

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

#(Intercept) -1.749e-01 1.659e-01 -1.054 0.291683

#OWNRENT -2.179e-01 7.686e-02 -2.835 0.004581 \*\*

#INCOME -2.424e-04 4.364e-05 -5.554 2.79e-08 \*\*\*

#INCPER -1.291e-05 3.318e-06 -3.890 0.000100 \*\*\*

#LOGSPEND -2.165e-01 2.848e-02 -7.601 2.95e-14 \*\*\*

#AGE -8.330e-03 3.774e-03 -2.207 0.027312 \*

#EXP\_INC 1.340e+00 3.467e-01 3.865 0.000111 \*\*\*

BIC(m1)

# 6401.586

roc1 <- pROC::roc(response = df2$DEFAULT, predictor = fitted(m1))

# Area under the curve: 0.6554

m2 <- glm(eval(paste(y, paste(x2, collapse = " + "), sep = " ~ ")), data = df2, family = binomial)

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

#(Intercept) -1.222e+00 9.076e-02 -13.459 < 2e-16 \*\*\*

#MAJORDRG 2.031e-01 6.921e-02 2.934 0.00335 \*\*

#MINORDRG 1.920e-01 4.784e-02 4.013 5.99e-05 \*\*\*

#INCOME -4.706e-04 3.919e-05 -12.007 < 2e-16 \*\*\*

BIC(m2)

# 6421.232

roc2 <- pROC::roc(response = df2$DEFAULT, predictor = fitted(m2))

# Area under the curve: 0.6429

Both AUC and BIC statistics seemed to favor the first model. However, is a 2% difference in AUC significant enough to infer a better model? Under the Null Hypothesis of no difference in AUC, three statistical tests were employed to assess the difference in AUC / ROC between two models.

set.seed(2019)

# REFERENCE:

# A METHOD OF COMPARING THE AREAS UNDER RECEIVER OPERATING CHARACTERISTIC CURVES DERIVED FROM THE SAME CASES

# HANLEY JA, MCNEIL BJ (1983)

pROC::roc.test(roc1, roc2, method = "bootstrap", boot.n = 500, progress = "none", paired = T)

# D = 1.7164, boot.n = 500, boot.stratified = 1, p-value = 0.0861

# REFERENCE:

# COMPARING THE AREAS UNDER TWO OR MORE CORRELATED RECEIVER OPERATING CHARACTERISTIC CURVES: A NONPARAMETRIC APPROACH

# DELONG ER, DELONG DM, CLARKE-PEARSON DL (1988)

pROC::roc.test(roc1, roc2, method = "delong", paired = T)

# Z = 1.7713, p-value = 0.0765

# REFERENCE

# A DISTRIBUTION-FREE PROCEDURE FOR COMPARING RECEIVER OPERATING CHARACTERISTIC CURVES FROM A PAIRED EXPERIMENT

# VENKATRAMAN ES, BEGG CB (1996)

pROC::roc.test(roc1, roc2, method = "venkatraman", boot.n = 500, progress = "none", paired = T)

# E = 277560, boot.n = 500, p-value = 0.074

Based upon the above output, there is no strong statistical evidence against the Null Hypothesis.

pscl::vuong(m1, m2)

# Vuong z-statistic H\_A p-value

#Raw 2.0963830 model1 > model2 0.018024

#AIC-corrected 1.8311449 model1 > model2 0.033539

#BIC-corrected 0.8684585 model1 > model2 0.192572

In addition, a Vuong test is also performed, supporting no difference between two models after corrected for the Schwarz penalty.