## Use Confusion Matrices as a performance evaluation tool. This would be perfect for evaluating a model’s predictions and whether or not they match actual outcomes. This is only suitable for any evaluation type questions you may come across.

Call**:**

glm**(**formula **=** obese **~** flipper\_length\_mm **\*** bill\_length\_mm **+** sex **+**

species, family **=** binomial**(**link **=** "logit"**)**, data **=** train\_data**)**

Coefficients**:**

Estimate Std. Error

**(**Intercept**)** 27.552190 78.862421

flipper\_length\_mm **-**0.207007 0.406718

bill\_length\_mm **-**1.384759 1.867541

sexmale 2.892290 1.094641

speciesChinstrap **-**4.739485 1.655085

speciesGentoo 2.640500 1.866750

flipper\_length\_mm**:**bill\_length\_mm 0.008385 0.009612

z value Pr**(>|**z**|)**

**(**Intercept**)** 0.349 0.72681

flipper\_length\_mm **-**0.509 0.61077

bill\_length\_mm **-**0.741 0.45840

sexmale 2.642 0.00824 **\*\***

speciesChinstrap **-**2.864 0.00419 **\*\***

speciesGentoo 1.414 0.15722

flipper\_length\_mm**:**bill\_length\_mm 0.872 0.38300

**---**

Signif. codes**:**

0 ‘**\*\*\***’ 0.001 ‘**\*\***’ 0.01 ‘**\***’ 0.05 ‘.’ 0.1 ‘ ’ 1

**(**Dispersion parameter **for** binomial family taken to be 1**)**

Null deviance**:** 323.002 on 232 degrees of freedom

Residual deviance**:** 98.718 on 226 degrees of freedom

AIC**:** 112.72

Number of Fisher Scoring iterations**:** 8

Analysis:

Flipper Length: For a 1mm increase in flipper length, the log-odds of obesity decrease by 0.207, holding other predictors constant (ceteris paribus).

0.813 (therefore a 1mm increase reduces odds of obesity by roughly 19%.)

However, as its standard error is 0.406718, which is not significant (as the p value is 0.61077), therefore the effect that the flipper length could have might be imprecise; therefore there’s a lack of confidence in saying that flipper length actually impacts obesity on its own.

Due to similar nature in the estimate and std. error variables for Bill Length, the interpretation is the same.

‘sexmale’: Since its estimate is 2.892290, compared to females, males have a higher log-odds of obesity, being increased by 2.892.

Therefore, males are inherently more likely to be obese. Mathematically you could calculate this by figuring out the odds ratio:

which means that males have 18-times higher odds of obesity than females. This is essentially true as bigger, fatter male penguins tend to have higher odds of mating with females to prolong their generations.

With the standard error of 1.094641, and with the P value being 0.00824 (therefore being significantly larger), there’s strong evidence to suggest that males are more prone to obesity than females, with a reasonably precise estimate.

Apply the same calculations and structure for the rest of the intercepts.

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 42 2

1 13 43

Accuracy **:** 0.85

95% CI : (0.7647, 0.9135)

No Information Rate **:** 0.55

P**-**Value **[**Acc **>** NIR**]** **:** 1.716e-10

Kappa **:** 0.7036

Mcnemar's Test P-Value : 0.009823

Sensitivity : 0.9556

Specificity : 0.7636

Pos Pred Value : 0.7679

Neg Pred Value : 0.9545

Prevalence : 0.4500

Detection Rate : 0.4300

Detection Prevalence : 0.5600

Balanced Accuracy : 0.8596

'Positive' Class : 1

**Analysis:**

The model accurately predicted the positive and negative class (42+43) 85 times out of 100.

The Cohen’s Kappa metric measures agreements between the predictions and actuals, adjusted for chance, ranging from 0 (no agreement beyond chance) to 1 (perfect agreement). Therefore, the figure of 0.7036 indicates a substantial level of agreement.

The McNemar’s Test P-Value tests if the model’s errors (FP vs. FN) are symmetric. A p-value < 0.05 (like 0.009823) suggests the errors are significantly asymmetric (e.g., more FN than FP).

**Sensitivity:**

* The model correctly identifies 95.56% of the actual positives.

**Specificity:**

* The model correctly identifies 76.36% of the actual negatives.

**Positive Predictive Value:**

* When the model predicts positive, it’s correct 76.79% of the time.

**Negative Predictive Value:**

* When the model predicts negative, it’s correct 95.45% of the time.

**Balanced Accuracy:**

* Balances performance across both classes (sensitivity and specificity); here it’s 85.96%.

**Class 1 (Positive):**

* The model excels at identifying positives (sensitivity = 95.56%), but it has more false negatives (13) than false positives (2), as reflected in the McNemar’s test.

**Class 0 (Negative):**

* Specificity (76.36%) is decent but lower than sensitivity, meaning it’s less reliable at identifying negatives.

The model is likely useful in scenarios where detecting positives (class 1) is critical (high sensitivity), though it occasionally misses positives (FN = 13) or mislabels negatives as positives (FP = 2).

Therefore, with these values we can suggest (as there is sufficient evidence to do so) that the model is strong, as it’s significantly better than randomly guessing (50%). The high Kappa confirms a good agreement beyond chance.

> cat("Improved Accuracy:", round(imp\_accuracy, 3),

+ "- Proportion correct;\n")

Improved Accuracy: 0.85 - Proportion correct;

> cat("Improved Sensitivity:", round(imp\_sensitivity, 3),

+ "- Detection of obese;\n")

Improved Sensitivity: 0.956 - Detection of obese;

> cat("Improved Specificity:", round(imp\_specificity, 3),

+ "- Detection of non-obese;\n")

Improved Specificity: 0.764 - Detection of non-obese;

> cat("Improved Precision:", round(imp\_precision, 3),

+ "- Reliability of obese predictions;\n")

Improved Precision: 0.768 - Reliability of obese predictions;

**Analysis:**

**Improved Accuracy: 0.85 - Proportion correct**

* Value: 0.85 (85%)
* Meaning: The proportion of all predictions (both obese and non-obese) that the improved model got correct. This matches the accuracy from your confusion matrix: (TP + TN) / Total = (43 + 42) / 100 = 0.85.
* Context: This is the overall success rate of the model.

**Improved Sensitivity: 0.956 - Detection of obese**

* Value: 0.956 (95.6%)
* Meaning: The proportion of actual obese individuals correctly identified by the model. This aligns with the sensitivity from your earlier metrics: TP / (TP + FN) = 43 / (43 + 13) ≈ 0.9556 (rounded to 0.956 here).
* Context: The model is highly effective at detecting obese individuals, missing very few cases (only 13 false negatives out of 56 obese cases).

**Improved Specificity: 0.764 - Detection of non-obese**

* Value: 0.764 (76.4%)
* Meaning: The proportion of actual non-obese individuals correctly identified by the model. This matches the specificity from earlier: TN / (TN + FP) = 42 / (42 + 2) ≈ 0.7636 (rounded to 0.764 here).
* Context: The model is moderately good at identifying non-obese individuals but makes more mistakes here (2 false positives out of 44 non-obese cases) compared to detecting obese individuals.

**Improved Precision: 0.768 - Reliability of obese predictions**

* Value: 0.768 (76.8%)
* Meaning: The proportion of predicted obese cases that are actually obese. This corresponds to the positive predictive value (PPV) from earlier: TP / (TP + FP) = 43 / (43 + 2) ≈ 0.7679 (rounded to 0.768 here).
* Context: When the model predicts someone is obese, it’s correct 76.8% of the time. The remaining 23.2% are false positives (non-obese individuals incorrectly labelled as obese).

Under each defined value includes a relatively short explanation for each component (Accuracy, Sensitivity, Specificity, Precision (according to the functions of the R library)).

The model is designed to classify individuals as obese or non-obese, likely for a medical or health-related purpose.

* Strength: It’s excellent at detecting obese individuals (sensitivity = 95.6%), which is critical if the goal is to identify as many obese cases as possible (e.g., for early intervention).
* Trade-off: It’s less reliable at ruling out obesity (specificity = 76.4%), meaning some non-obese individuals might be misclassified as obese. Precision (76.8%) also indicates that about 1 in 4 "obese" predictions is incorrect.