Applied Statistics for Data Scientists with R

Class 19: GLM and Logistic Regression

What is GLM?

- GLM (**Generalized Linear Model**) is a flexible generalization of ordinary linear regression that allows for response variables to have error distributions other than a normal distribution.
- If the residuals (resulting from the response variable) do not follow normal distribution, then such method helps.

Odds

$$Odds = \frac{Probability\ of\ Success}{Probability\ of\ Failure} = \frac{P(Success)}{1 - P(Success)}$$

- So, if a student has 0.75 probability of passing an exam, then the probability of failure is 0.25
- Hence, $odds = \frac{0.75}{0.25} = 3.$
- It means the student is three times more likely to pass than fail.

Logistic Regression

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$\log\left(\frac{P(Survived)}{1-P(Survived)}\right) = \beta_0 + \beta_1 Pclass + \beta_2 Sex + \beta_3 Age + \beta_4 Fare$$

Example: Predicting Chance of Survival

- Pclass (-1.27): Higher-class passengers had better survival odds. Each increase in class (from 1st to 2nd or 2nd to 3rd) decreases the log-odds of survival by 1.27 (lower class = lower survival chance).
- SexFemale (2.52): Being female significantly increases survival odds. The log-odds increase by 2.52 if the passenger is female compared to a male. This means women were much more likely to survive.
- Age (-0.037): Older passengers had a lower survival probability. A 1-year increase in age decreases log-odds by 0.037, meaning younger people had slightly better survival odds.
- Fare(0.00054): Fare has a very small positive effect on survival, but it's not statistically significant (p = 0.805), meaning we can't conclude that fare significantly impacted survival.

Performance Evaluation

Predicted	Reference: Event	Reference: No Event
Event	A	B
No Event	C	D

The formulas used here are:

$$Sensitivity = \frac{A}{A+C}$$

$$Specificity = \frac{D}{B+D}$$

$$Prevalence = \frac{A+C}{A+B+C+D}$$

$$PPV = \frac{sensitivity \times prevalence}{((sensitivity \times prevalence) + ((1 - specificity) \times (1 - prevalence))}$$

$$NPV = \frac{specificity \times (1 - prevalence)}{((1 - sensitivity) \times prevalence) + ((specificity) \times (1 - prevalence))}$$

$$Detection\ Rate = \frac{A}{A+B+C+D}$$

$$Detection\ Prevalence = \frac{A+B}{A+B+C+D}$$

$$Balanced\ Accuracy = (sensitivity + specificity)/2$$

$$Precision = \frac{A}{A+B}$$

$$Recall = \frac{A}{A+C}$$

$$F1 = \frac{(1+\beta^2) \times precision \times recall}{(\beta^2 \times precision) + recall}$$

From caret package: https://topepo.github.io/caret/measuring-performance.html

Confusion Matrix & Related KPIs

```
          Reference
Prediction   0   1
         0 357  81
         1  67 209
```

	Actual Not Survived	Actual Survived	Total
Predicted Not Survived	TN = 357	FN = 81	424
Predicted Survived	FP = 67	TP = 209	290
Total	438	276	714

- True Negatives (TN = 357): Model correctly predicted non-survivors.
- False Positives (FP = 81): Model incorrectly predicted some non-survivors as survivors.
- False Negatives (FN = 67): Model missed predicting some actual survivors.
- True Positives (TP = 209): Model correctly predicted survivors.

Confusion Matrix & Related KPIs

	Actual Not Survived	Actual Survived	Total
Predicted Not Survived	TN = 357	FN = 81	438
Predicted Survived	FP = 67	TP = 209	276
Total	424	290	714

$$\text{Accuracy} = \frac{TP + TN}{Total}$$

- model correctly classifies survival status about 79% of the time.

$$\text{NIR} = \frac{\max(\text{Class Counts})}{\text{Total}}$$

- If we blindly guessed the majority class (non-survivors), we'd be right 59.38% of the time.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- The model correctly identifies 84.2% of actual survivors. Also called recall.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- The model correctly identifies 72.07% of non-survivors.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- 81.51% of passengers predicted as "Survived" actually survived.

Confusion Matrix & Related KPIs

	Actual Not Survived	Actual Survived	Total
Predicted Not Survived	TN = 357	FN = 81	438
Predicted Survived	FP = 67	TP = 209	276
Total	424	290	714

McNemar's Test (p = 0.2853)

- Tests if false positives and false negatives occur at the same rate.
- $p > 0.05$: No significant difference between misclassifications.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

- Average of Sensitivity (detecting survivors) and Specificity (detecting non-survivors).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- F1 = 82.83%: A harmonic mean of precision and recall.
- Balances both False Positives & False Negatives.

$$\text{Prevalence} = \frac{\text{Total Positive Cases}}{\text{Total Samples}}$$

- 59.38% of passengers did NOT survive

Kappa Statistic

- The Kappa statistic measures how well a classification model performs beyond random chance. Unlike accuracy, which can be misleading when classes are imbalanced, Kappa accounts for chance agreement.
- Ranges from -1 to +1
  - 1: Perfect agreement
  - > 0.7 → Good agreement.
  - 0.5 - 0.6 → Moderate agreement.
  - 0: No better than chance
  - Negative: Worse than random guessing