MODULE 5

Logistic regression
A technique that models a categorical dependent variable y, based on one or more independent variables x.

Binomial logistic regression
A technique that models the probability of an observation falling into one of two categories, based on one or more independent variables.

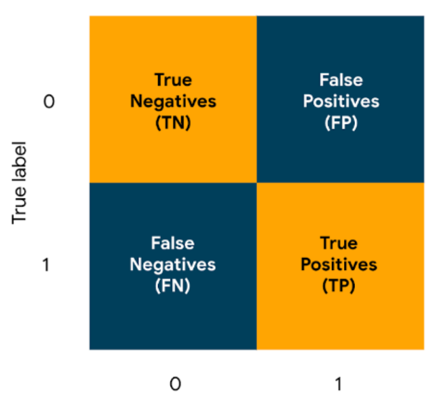
Assumptions:

- **Linearity**
  Linear relationship between each x variable and the logit of the probability that y equals 1
  - $Odds\ =\ \frac{p}{1-p}$
  - **Logit (log-odds)** - Logarithm of the odds
    - $logit(p)\ =\ log(\frac{p}{1-p})$
    - Common link function used to linearly relate the x variables to the probability of y
    - Logit in terms of x variables
    - $logit(p)\ =\ log(\frac{p}{1-p}) = \beta_0\ +\ \beta_1 X_1\ +\ \beta_2 X_2 + \ldots + \beta_n X_n\ \ where\ p\ =\ P(Y = 1)$
    - For every 1 unit increase in x, we can expect that the y odd to increase by $(1 - e^{\beta_1})$

  - **Maximum likelihood estimation (MLE)**
    - A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data
    - The best logistic regression model estimates the set of beta coefficients that maximizes the likelihood of observing all of the sample data

- **Independent observations**
  - $P(A\ AND\ B)\ =\ (P(A)\ *\ P(B)$
- **No multicollinearity**
- **No extreme outliers**
  - Transform or adjust variables
  - Remove the outliers

Confusion matrix – A graphical representation of how accurate a classifier is a predicting the labels for a categorical variable



Accuracy – The proportion of data points that were correctly categorized

$$Accuracy = \frac{No.of\ correct\ predictions}{No.of\ total\ predictions}$$

Precision – Proportion of positive predictions that were true positives

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall – Proportion of positive the model was able to identify correctly

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

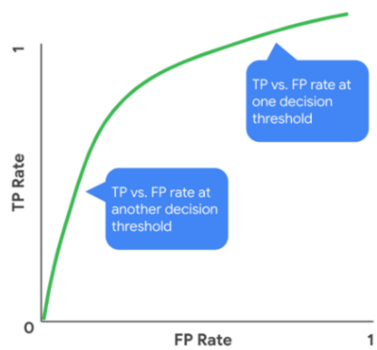
F1 score – Harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Accuracy, precision, recall and F1 score are useful for measuring unbalanced classes

ROC (receiver operating characteristic curve)

- Helps in visualizing the performance of a logistic regression classifier
- Classification threshold is a cutoff for differentiating the positive class from the negative class
- In an ideal model, the threshold exists at which TPR is high and FPR is low (curve hugs top left corner)



- **True Positive Rate** (equivalent to recall)

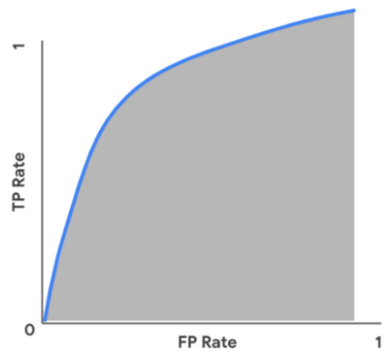
$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **False Positive Rate**

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

AUC (Area Under the ROC Curve)

- Provides an aggregate measure of performance across all possible classification thresholds
- Ranges from 0.0 to 1.0
- Model with 100% wrong predictions have AUC of 0.0 and 100% right have AUC of 1.0
- AUC < 0.5 indicates that the model performs worse than a random classifier
- AUC > 0.5 indicates that the model performs better than a random classifier
- AUC is the area of the shaded region


1
TP Rate
0
FP Rate
1