LOGISTICS REGRESSION – Binary Classifications

Sigmoid Curve

F(x) = 1/1+e-x 🡪 p = 1/1+e-x

Max. Likelihood

P = 0 [0,1]

P = 1 x>0

X = b0+b1x1+b2x2

F(y) = p = 1/1+e-( b0+b1x1+b2x2)

Ln(p/1-p) = b0+b1x1+b2x2+………. Log of Odds 🡪0/1

P 🡪 [0,1]

* Choose the cut-off value wisely.

Advantages

Easy to implement

Disadvantages

Issue with overfitting

Only used for Binary Classification

Cost Function: Cross entropy function

-(ylogp + (1 – y)log(1-p))

Y = 0 –(1-y)log(1-p)

Y = 1 -ylogp

Metrics: - Confusion Matrix (depends on cut -off)🡪 Similar to Pivot table

* Comparison between Actual vs predicted y values.

It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting.

This is how it looks like:

Graphical user interface, application

Description automatically generated

**Accuracy**

Graphical user interface, text

Description automatically generated

**Sensitivity** =TP / TP + FN

**Specificity =** TN / TN + FP

**Specificity and sensitivity**

Table

Description automatically generated with low confidence

**Recall =** TP / TP + FN

**Precision =** TP / TP + FP

**Recall and Precision**

Text

Description automatically generated

**Formula for recall and sensitivity is same.**

**Higher is the ROC value good is the model.**

**Area under the curve**

**ROC (Receiver Operating Characteristic) =** Sensitivity vs 1-Spectivity

The plot of ‘True Positive Rate’ (Sensitivity/Recall) against the ‘False Positive Rate’ (1-Specificity) at different classification thresholds. The area under the ROC curve (AUC) measures the entire two-dimensional area underneath the curve. It is a measure of how well a parameter can distinguish between two diagnostic groups. Often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1.

Chart

Description automatically generated

Critical Points

If TPR = FPR =0 🡪 Your classifier predicts all the instances to be -ve

If TPR=FPR=1 🡪 Your classifier predicts all the instances to be +ve

If TPR =1 and FPR=0 🡪 Ideal class in ROC.

PROS – Simple Easy to implement

CONS – Over -Fitting, Binary class, Outliers, missing values

We need to convert categorical column which has multiple levels into binary format, by doing dummy variable creation it will expand the columns into multiple columns.

A picture containing table

Description automatically generated

Feature Elimination -RFE – Recursive Feature Elimination

It will help us to take columns having higher impact on column

Graphical user interface, text, application

Description automatically generated with medium confidence

Logistics regression Steps

1] Dummy variable 🡪 Feature Engineering 🡪 One hot Encoding

2] Recursive Feature Elimination 🡪 to remove unwanted columns

3] Then take route of VIF and P value

4] Cut-off Tuning (0, 0.1, 0.2,0. 3.0.9)

5] Accuracy, Sensitivity, Specificity

6] Test dataset

7] GLM – Generalized Linear Model

8] Make sure the model is simple and generic

Disadvantage

Can’t be used f or multi linear